IEEE MICRO

DECEMBER 1992

Chips, Systems, Software, and Applications

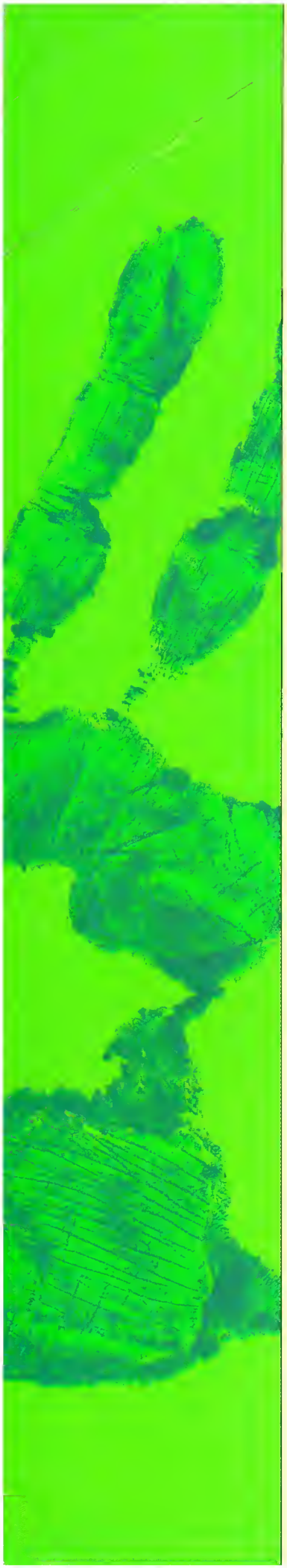
Object Recognition

& Other Special-Purpose DSPs

Also: Associative Processors/Memories, Part 2

IEEE COMPUTER SOCIETY

THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.



IEEE MICRO

Published by the IEEE Computer Society

Volume 12 Number 6 December 1992

FEATURES

THEME ARTICLES

ASSOCIATIVE SYSTEMS, PART 2

SPECIAL FEATURE

Cover: Chris Martin, Design and Direction

Circulation: *IEEE Micro* (ISSN 0272-1732) is published bimonthly by the IEEE Computer Society, PO Box 3014, Los Alamitos, CA 90720-1264; IEEE Computer Society Headquarters, 1730 Massachusetts Ave., NW, Washington, DC 20036-1903; IEEE Headquarters, 345 East 47th St., New York, NY 10017. Annual subscription: $23 in addition to IEEE Computer Society or any other IEEE society member dues; $39 for members of other technical organizations. This journal is also available in microfiche form.
Postmaster: Send address changes and undelivered copies to *IEEE Micro*, PO Box 3014, Los Alamitos, CA 90720-1264. Second-class postage is paid at New York, NY, and at additional mailing offices. Canadian GST#125634188.

DEPARTMENTS

Call for Articles, p. 23; Reader Service cards, p. 64A/B; Advertiser/Product Index/ Moving Coupon, p.96; Computer Society information, cover 3

IEEE Computer Society
PO Box 3014, Los Alamitos, CA 90720-1264
(714) 821–8380

* Submit six copies of all articles and special-issue proposals to Dante Del Corso, Dipartimento di Elettronica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129 Torino, Italy; phone +39-11-556-4044; Compmail: d.delcorso; Internet: delcorso@polito.it; Bitnet: delcorso@itopoli
or
** Ashis Khan, Mips Computer Systems, Inc., 950 DeGuigne Drive, Sunnyvale, CA 94086; Internet: ashis@mips.com.
Author guidelines for article submission are available from the West Coast office (address at left).

From the Editor-in-Chief

Letters and special themes

SHOULD WE GO BACK to split articles? When articles were split, we received several letters with complaints on splitting and some other proposals. After returning to continuous articles, we received neither letters with complaints nor other suggestions.

We appreciate any comment or proposal you send to us and, when possible, we implement them. We normally provide space for you to add a few sentences on how we can improve *IEEE Micro* as you circle numbers in the product part of the Reader Service Card. In this issue and the last issue, however, we added survey cards to elicit detailed comments on how we are serving you. We will use your comments to complete information already collected through sample telephone interviews.

Please help us to improve *Micro* (and serve you better). We also welcome any comment or proposal besides the predefined questions.

This issue features the theme of special signal processors, which is described more fully in the guest editors' introduction. We also complete the articles that could not fit in an earlier special issue.

Last June, *Micro* presented a special issue on the re-emerging field of associative (a synonym for content-addressable) systems. Part 2 of this theme deals with approaches that extend the normal CAM structure by powerful logic components to achieve real associative processor systems. These systems not only retrieve data but also permit their processing to be organized according to some properties of the data.

The first article, by K.E. Grosspietsch and R. Reetz, discusses an experimental architecture of an associative system with several innovative features, for example, inclusion of some processing logic within the bit cells.

The second article, by C.D. Stormon, N.B. Troullinos, E.M. Saleh, A.V. Chavan, M.R. Brule, and J.V. Oldfield, describes the design and application of an associative processor chip. This chip is a CAM architecture augmented by additional processing logic for each memory word.

Let us know how you feel about these special themes and the others we've featured this year; *IEEE Micro* is your magazine and should reflect your interests.

In the mailbag

(LK: liked; DLK: disliked; LTS: like to see)

December 1991

LK: Send some catalogs and magazines....—M.R.S., Isfahan, Iran (This is done through the product information part of the Reader Service Card; fill the card out and send it in to us.—D.D.C.)

LK: The information on computer boards and cards.—J.R., Grenoble, France

February 1992

LK: Scalable Coherent Interface [and] neural network classifier [articles]; LTS: application-specific architectures—A.S.M., Secunderabad, India

0272-1732/92/1200-0002$03.00 © 1992 IEEE

Quality standards

Stephen L. Diamond

SunSoft, Inc.

(415) 336-4190

fax (415) 336-4477

steve.diamond@eng.sun.com

With this issue, I'm honored to join the *IEEE Micro* Editorial Board as editor of Micro Standards. As you may know, Carl Warren, the former editor, passed away in July after a long illness. I knew Carl for many years, both as a colleague in the computer industry and through his participation in IEEE standards, in particular his work in developing software language standards for Forth and Pilot. Carl was a strong advocate of IEEE standards activities, and he will be sorely missed.

Let me briefly introduce myself. I've been involved in IEEE standards activities for more than a decade. I currently chair the IEEE Computer Society Microprocessor Standards Committee, which develops standards for microprocessor and microcomputer systems. I serve as a member of the US Technical Advisory Group to ISO/IEC JTC1 Subcommittee 26 (microprocessor systems). I'm also director of standards at SunSoft.

I'd like to begin my tenure as editor with a discussion of some general concepts related to standards. Everyone has a definition of standards and standardization, and no two exactly match. People describe good and bad standards—and then begin to argue about what is meant by "good" or "bad." I'll start by quickly examining exactly what standards are and move forward from there.

Standards as tools

In my view, we should see standards as tools—and, as tools, they are intrinsically neither good nor bad. What's important is the fit between the tool and the task. A particular wrench might be a poor implementation of "wrenchness," but the significant issue is whether a wrench is useful in a particular context. It might be a very useful tool for removing a car's exhaust manifold, but a singularly useless one for removing a screw. Likewise, the IEEE floating-point standard, IEEE Std. 754-1991, is particularly useful in the world of microcomputers and not nearly as useful in the world of mainframes.

This argument becomes complicated when one asks exactly who the audience is for these tools. Many Standards Developing Organizations (SDOs) implicitly assume that the user of standards also uses the technical document produced. The real user may be someone who has no interest in the technical content of the standard, but rather wants to buy something that is "standard." This is where the problem of defining good and bad standards becomes really obscured, because there is no metric by which to judge perceived goodness or badness.

G.K. Chesterton might have been talking about standards when he said, "The word 'good' has many meanings. For example, if a man were to shoot his grandmother at a distance of 500 yards, I should call him a good shot, but not necessarily a good man." Whether Std. 754 is useful in a particular marketplace does not decide whether it is a good or bad standard. Indeed, many believe that 754 is a good standard because it is relatively clear, concise, and unambiguous. My measure here of "goodness" is that the document is easily implementable by a technical person producing software or hardware who wants to implement floating point in a fashion that is good for microcomputers.

On the other hand, I might think the standard was bad if I used a brand of mainframes that would not share data with PCs that were being incorporated into my network because they used 754. The definition of "badness" here is one of failed expectation.

This, then, is a major conundrum, caused by

0272-1732/92/1200-0003$03.00 © 1992 IEEE

		Form: Good	Form: Bad
Function (for a particular purpose)	High	1	2
	Low	3	4

Figure 1. A simple matrix showing the relationship between form and function for a standard.

the use of ambiguous terms like "good" and "bad," complicated by differences in judgmental perceptions (creator versus implementor versus economic user). To avoid this, let's look at two critical dimensions of a standard, the form and function.

Form and function

We can define a standard in terms of its "form"—that is, based on criteria such as clarity, conciseness, level of ambiguity, timeliness, process, and openness (CCATOP). Such criteria rate how well a particular standard matches the form of an "ideal" standard. Many creators of standards judge their work against this criteria. This is, however, only half the equation for judging a standard.

The concept of a standard's "function" is more difficult to define than its form. The focus of standards in the IEEE Computer Society, as defined in its *Policies and Procedures Manual*, is on the creation of "broadly accepted, sound, and technically excellent standards that will advance the theory and practice of computer science and engineering." One of the functions of standards is to increase the general competence of engineers worldwide and the other is to be broadly accepted. Note that the functional requirements of a standard are the province and concern of the standard's users rather than its creators.

Therefore, I believe we can measure the function of an IEEE standard by its technical excellence and by whether it (and the solution it proffers) is accepted by the industry, which comprises technologists designing from the standard, purchasing agents specifying compliance to the standard, and the large number of end users willing to pay for products embodying the standard's technical attributes. This wide range of users makes the function of a standard very difficult to quantify, because a standard's utility becomes a derivative of its context. A design engineer's contextual setting is substantially different from that of a purchasing agent. One wants to build a product to sell, and the other wants to buy a product that performs a job or process.

It is difficult to constantly refer to the form and function of a standard. To make these terms easier to use, I've developed a simple matrix, as shown in Figure 1.

The standards in quadrant 1 are those that have both form and function (within a particular context) that meet the highest expectations. Standards in quadrant 2 may be poorly written (or lack some of the other CCATOP requirements), but they meet the functional requirements. Those in quadrant 3 are less desirable, because they may have all the CCATOP functions, but for some reason are not really useful. Finally, those in quadrant 4 fill no function, are largely incomprehensible, and can be called low-quality standards.

Defining quality standards

Let's turn to the term "quality," defined as "fitness for purpose," which suits the discussion well. We can judge the quality of IEEE standards by how well they embody CCATOP principles and whether they are widely accepted and technically excellent. A quality standard meets these criteria—that is, it expands the discipline's general knowledge or it makes the technology more useful to users of the standard, whoever they are. This is a combination of form and function: the quality standard is fit for the purpose for which it was designed.

Using this definition, one set of users of the standards are designers of "things" who need to understand the technology that drives their profession. The measure of success for these engineers is whether they can make their technical discipline accessible to an ever-wider range of people who can apply the technology to anything from designing an airplane, to building a house, to balancing a checkbook. At the same time, buyers of products designed around these standards (end users) are also users of a standard, although they may be classified as secondary users.

To return to our tool analogy, the designer of a hammer must understand that carpenters judge hammers by one criterion ("hammerness"), but that they buy hammers for one reason—to use them. If the job instead calls for staples, the hammer—even though it might have good "hammerness"—will lose out to a power stapler, because the hammer is the wrong tool for the desired function. It is the same with standards. If the users do not believe in them and do not use them, the standards have failed.

Achieving high quality

The key to success for this effort is to integrate the creators' needs with those of the users. The creators must clearly understand who the potential users of their standards are, and they must communicate with their users to discover their needs. For their part, the users must know exactly they want. The three parties (creators, designers, and end users) involved in the standardization effort must integrate their requirements and be willing to work together.

I believe this form of integration is possible, but will require really hard work and a challenge to old and dearly held ideas. It will often require the rethinking and reasking of questions, some of which may be uncomfortable or downright scary. The standardization creators will have to talk to working design engineers who may deride

the work of a committee—or the design engineers may have to accept that a standardized design could be better than their work. End users may not see the benefit of good technology because what they have is fine, or they may demand something impossible to achieve. All three parties bear a heavy burden to make the process function as it can—and should—work.

The most difficult thing about working together is that we must reevaluate the questions we are trying to solve in light of an entirely new external environment. What we took for granted in 1989 and 1990 is no longer valid; users have changed, the information technology industry has changed, and the political and economic climates in which standardization takes place have changed.

I think a high-quality standard is one that advances the study, understanding, and utility of standards as a tool. This definition allows us to separate the standard into form and function, and then to judge the standard against several criteria. Finally, it allows a clearer definition of where we want to go with these things called standards, which will permit us to plan how to achieve our goals—an idea I believe Carl Warren would have liked.

Richard H. Stern

Oblon, Spivak, McClelland, Maier & Neustadt, P.C.

1755 Jefferson Davis Highway, Suite 400

Arlington, VA 22202

Unobserved Demise of Exhaustion Doctrine

For over a century, the exhaustion doctrine, or first-sale rule, was a fixture of US patent law. It said that a patent owner could not prevent customers from freely using and reusing patented products purchased from the patent owner or its authorized distributors and licensees. But recently, the US Court of Appeals for the Federal Circuit (CAFC) overruled this doctrine—a remarkable decision that went largely unnoticed. The new rule has wide ramifications for manufacturers and their customers.

Under the exhaustion doctrine, a patentee could, in many circumstances, give a manufacturer a limited license to exploit a patent only in a particular field. For example, the owner of an amplifier patent might license A to make and sell large amplifiers for theaters and license B to make and sell small amplifiers for radios and phonograph machines. In such a case, B would infringe on the patent if it made and sold theater amplifier systems. But if the same patentee sold amplifiers to customers, it could not prevent them from using the amplifiers in theaters or wherever else they pleased, or from reselling the amplifiers to others for any use that they might please to make (unless the use infringed some other, unrelated patent).

The term "exhaustion doctrine" means that the patentee's sale of the product "exhausts" the patent monopoly. This rule was so clear that in 1926 the US Supreme Court observed:

> It is well settled ... that where a patentee makes the patented article and sells it, he can exercise no further control over what the purchaser may wish to do with the article after his purchase. It has passed beyond the scope of the patentee's rights.

More recently, at the time Congress passed the Semiconductor Chip Protection Act of 1984–which, like the US Copyright Act, contains an express recitation of the exhaustion doctrine–the House Report accompanying the bill stated:

> Section 906(b) carries over to mask works the "exhaustion of monopoly rights" and "first sale" doctrine of 17 U.S.C. § 109(a) and many years of case law. As in the case of copyrighted products, the owner of a mask work has no right to try to exercise "remote control" over the pricing or other business conduct of its semiconductor chip customers, once

0272-1732/92/1200-0005$03.00 © 1992 IEEE

> the semiconductor chips have passed into their hands. Except where the Congress expressly orders otherwise, the exhaustion of any rights by the first authorized sale is a basic tenet of our intellectual property law.

In a 1964 extension of the rule, the Supreme Court held that a logical consequence of the exhaustion doctrine was that purchasers of patented machines had a right to make enhancements without becoming liable as infringers for doing so. Thus a purchaser of a patented machine had the right to modify it to change the size of products it made or to increase its throughput. In the software field, this doctrine led to a similar right of purchasers of copies of copyrighted computer programs to modify them to add features or to port the programs to other platforms without becoming liable as copyright infringers. In general, Europe and Japan also follow the exhaustion doctrine and have done so for many years.

But this September, without drawing much attention, the CAFC overruled this whole body of patent law. In *Mallinckrodt, Inc.* v. *Medipart, Inc.*[1] the plaintiff, Mallinckrodt, owned a patent on a device that dispenses a radioactive mist used in making certain diagnostic X rays and then traps the mist. Mallinckrodt sold the device to hospitals for about $40 to $50 and labeled it "single-use only." After using the device, a hospital would send it to a hazardous-waste removal site. Because the device itself costs approximately $10 to make, we may infer that most of the purchase price represented the value of the patent (or the patented technology). This fact inspired the defendant, Medipart, to go into the recycling business. For $20, Medipart would clean a hospital's device, put in a new filter, subject the device to gamma radiation to kill germs, and return the device to the hospital for reuse.

Because the recycler and hospitals were defying Mallinckrodt's "single-use only" labels and were cutting into Mallinckrodt's profits, Mallinckrodt sued for patent infringement. Medipart asked the district court to dismiss the patent infringement claim because of the exhaustion doctrine, and the court did so. (The district court also ordered Mallinckrodt to stop sending notices to hospitals warning them not to reuse the devices.) Mallinckrodt appealed, and the CAFC reversed.

The court held that, despite what the Supreme Court had said at various times, the exhaustion doctrine should apply only to cases in which the patentee seeks to impose price-fixing or tie-ins on its customers. Otherwise, patentees are free to limit their customers' use of products, as long as the restraints do not hinder competition enough to qualify as antitrust violations. The appeals court returned the case to the district court for a more comprehensive analysis of the competitive effects of the restriction and its general "reasonableness."

Medipart then threw in the towel and settled the case by agreeing to stop recycling the devices. The CAFC is the only court of appeals in the US that decides appeals from patent infringement trials. Furthermore, once a three-judge appeals panel of the court rules on a legal issue, all subsequent court panels must consider that ruling a binding precedent. The only way out of the first panel's ruling is for the CAFC judges to sit *en banc* and reverse the precedent by a majority of the whole court. Another way out is for the Supreme Court to reverse a later ruling of the CAFC that followed the given precedent, but the Supreme Court almost never entertains an appeal of a patent case from the CAFC. Therefore, we may expect the court's remarkable overruling of a century of Supreme Court decisional law to be binding for the foreseeable future.

The decision raises a number of questions. One is whether the CAFC's decision is legally supported. The ruling is insupportable and incredible; it suggests that something may be going wrong in the system because this is not an isolated instance of weird judicial action. However, I do not know what can be done about it.

Patentees are free to limit their customers' use of products, as long as the restraints do not hinder competition enough to be antitrust violations.

Whether or not the decision is legally sound, does it make good sense from a policy standpoint? It is hard to determine if a different legal rule would make more sense and have better results. For one thing, we are not writing on a clean slate. A great deal of business expectation, and business and marketing strategy, rest on what appeared to be settled law. No sensible adviser in this field would have expected a patentee to be able to limit what customers do with patented products that the patentee directly or indirectly sells them. It is uncertain how far the new rule goes, since it comes out of nowhere. But let us put all that aside.

The thrust of the new rule is to increase patent owners' economic power and, presumably, their ability to devise ways to extract revenue from their products and the technology those

products embody. By the same token, the rule takes power and/or revenue away from customers. This is illustrated by the obvious consequences of the court's decision on the plaintiff and defendant involved. Mallinckrodt will now earn an additional $20 or $30 for each unit that the recycler would have handled. The hospitals will pay an additional $20 for each unit that they previously could have recycled, and presumably this will in turn affect what they charge patients, insurers, Medicare, etc. Finally, Medipart will stop making its $10 or so per unit. (There may also be ecological and other impacts.)

Is all this good or bad? Mallinckrodt would say the added revenue encourages it to become more inventive and innovative, and that this incentive promotes technological progress. Presumably, Mallinckrodt has economic or technological reasons for not recycling the devices. Perhaps the reasons are sound, but the marketplace said it wanted recycling. There is no basis for presuming that the patentee's decision against recycling is better for all concerned than the marketplace's contrary decision.

The new rule affects much more than hospitals and medical devices. For example, many laser printer manufacturers are unhappy with toner refillers. The latter are companies that take empty, used toner cartridges; drill holes in them; fill them with toner; close the hole; and charge half the price of new, "officially" filled toner cartridges. What if laser printer manufacturers start selling toner cartridges labeled "for one use only"?

More broadly, the decision will affect how manufacturers address niche or differentiated markets. The value of the same technology may differ in different markets. A Motorola 68030 or Intel 80386 microprocessor chip, for example, may have different comparative advantage and value in these different end uses: personal computer, workstation, arcade video-game machine, home video-game console, microwave oven, and automobile. In some of these niches, a much cheaper Z80 chip may be just as good; in others, nothing else is as good; in still others, another chip may be much better, functionally. Accordingly, the maximum price that a 68030 chip can command should vary from niche to niche. Yet, if Motorola customers are free to use the chips as they please, or to resell them across markets, Motorola cannot maximize revenue by charging prices commensurate with the technology's value to the particular user. A workstation customer who would otherwise pay a high price can buy a chip from the microwave oven customer who pays a low price—for the same reason hospitals dealt with the recycler in the Mallinckrodt case. (They liked paying $20 better than paying $40.)

What if Motorola could sell a 68030 chip "for microwave oven use only" or "for home video use only"—and collect patent infringement damages from those who flouted the restriction? Would that be a better or worse way to run the semiconductor chip business? (Some of this goes on already: Some chips labeled "25 MHz" can actually run at 33 MHz but sell at the lower 25-MHz price. But right now, you don't get sued for patent infringement if you run a "25-MHz" chip at 33 MHz. Just wait.)

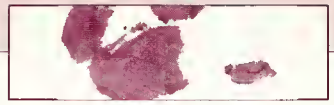
The argument in favor of expanding intellectual-property-law protections to how users use products embodying legally protected technology is largely that it enables manufacturers to achieve economies of scale and learning curve that otherwise could not be realized. (That is also what labeling some 33-MHz chips as 25-MHz chips does, but somewhat less effectively.) The other side of the argument is that it is too intrusive to have the law do this. Patent-law discipline (a species of government interference) now applies to many firms in the total market that previously were able to lead quiet lives without worrying about patent infringement matters and the significant transaction costs they involve (such as legal fees).

In addition, the patent laws embody a carefully negotiated bargain between the public and inventors. The public trades a limited amount of monopoly power and economic reward for increased disclosure of inventions. The century-old exhaustion doctrine is an element of the bargain. If the public ought to pay more, perhaps because we need more inventive incentives, who should decide that now—Congress or the CAFC?

There is no easy way to decide which argument has greater merit. The factors on the different sides are incommensurate—apples versus oranges. Good arguments can be made either way. In any event, a significant change in how patent law affects industry has sneaked up on us and is still largely unnoticed.

Reference

1. 24 U.S.P.Q.2d 1173 (Fed. Cir., Sept. 24, 1992).



Guest Editors' Introduction

Transforming the World of Digital Signal Processing

John L. Schmalzel

Parimal A. Patel

University of Texas at San Antonio

How far along are we in a paradigm shift that is likely to dramatically transform the world of digital signal processing? How will the architectural building blocks of DSP change?

Questions such as these will remain topical into the foreseeable future as progress in DSP techniques and architectures continues to transform the state of the art. DSP applications, in fact, drive integrated circuit technology in several important areas: architectures for special-purpose digital signal processors, high-speed data conversion devices, and others. Improvements in DSP components occur as a natural by-product of integrated circuit evolution, enhancing functional density and increasing switching speed. Other advances, such as novel techniques for implementing DSP algorithms and functions, assume increasing importance as designers develop fundamentally new approaches to solving traditional DSP problems.

We selected four broad themes for this special issue on DSP to bring you a cross-section of articles. Each article focuses primarily on one of the areas.

- **Recent information about digital signal processors.** Designers continue to improve DSP system architectures and, by adding resources to core functions, increase throughput. These improvements mirror general advances made in digital system architectures. In particular, reduced instruction-set computer (RISC) architecture is an area of active development. In fact, DSP architectures may be argued as the precursors to current RISCs in that DSP architectures are characterized by relatively limited instruction sets optimized to a class of applications.

  Michael Smith looks at developments in DSP architectures and asks a fundamental question, Are there significant differences between DSP and RISC architectures? He examines some typical DSP operations and compares their requirements to what is available in representative RISC architectures. This article also provides a good introduction to fundamental concepts of DSP architectures for those readers unfamiliar with the area.
- **Mixed analog/digital processors.** A basic feature of DSP applications is the need to use both analog and digital technologies in the system. As a minimum, analog techniques normally support the conversion processes needed to get analog signals into and out of the DSP system; that is, analog-to-digital and digital-to-analog conversion. Maximum system integration, "one-chip" solutions require that analog and digital domains be merged

0272-1732/92/1200-0008$03.00 © 1992 IEEE

on chip. Further advantages can accrue since reduction in physical dimensions can lead to other benefits such as improved signal-to-noise ratios from fewer interconnects and thus less interference.

Parimal Patel and coauthors report on development of a mixed-mode integrated circuit designed for biomedical DSP that integrates a 14-bit A/D with associated signal processing architecture elements.

- **Neural network techniques.** Alternatives to classical DSP approaches such as artificial neural networks and fuzzy logic have long been proposed and are now entering the mainstream. This becomes particularly evident when we view components recently introduced into the marketplace. Advantages of neural networks include freedom from precise system modeling. Instead, a training procedure provides the adaptation of a general network topology to an application.

  Jeff Brauch and his colleagues describe an application of a neural network integrated circuit for processing acoustic impact signatures. They used the Intel 80170NX, which consists of neuron elements, synapses, and input arrays. A unique feature of this chip is its external interface that allows it to operate directly on analog signals.
- **Other techniques for signal processing.** Classical DSP systems operate on discrete-time sequences obtained by sampling continuous-time signals. Examples of well-known techniques include spectral estimation using fast Fourier transforms and filtering using finite or infinite impulse response (FIR, IIR) filters. In addition to the neural network techniques just identified, designers seek other alternatives for signal processing.

  Jin Luo and his coauthors describe a novel approach for pattern recognition that is based on fundamentally analog techniques for perfoming DSP. Using a VLSI implementation of switch and resistive elements, their device solves a difficult pattern recognition problem.

Many of the authors speculate on future directions for their areas of effort. In selecting these articles for this issue, we neither attempted to clearly define the next generation of DSP nor cast aspersions on classical DSP approaches. However, we believe a number of issues arise after reading these articles, and these questions—like the two we posed in the opening paragraph—will occupy our efforts for some time. How will classical approaches to DSP continue to develop in the near future, and how will new techniques emerge to compete with them? When will new techniques shift the center of mass away from deterministic methods (such as computing fast Fourier transfonms) to stochastic and other nonlinear techniques (for example, neural networks)?

The world of DSP has been an interesting one, and it appears destined to remain so.

John L. Schmalzel is an associate professor of electrical engineering at the University of Texas at San Antonio. His teaching and research activities focus on instrumentation design, particularly microprocessor-based biomedical systems. His current work focuses on event recognition in long-term monitoring applications using fuzzy logic.

Schmalzel received his undergraduate and graduate electrical engineering degrees from Kansas State University. He is a member of the IEEE and a number of societies including the Computer Society. He serves on the Editorial Board of *IEEE Micro*.

Parimal A. Patel is an assistant professor of electrical engineering at the same university. His teaching and research interests include application-specific integrated electrode design, ASIC DSP processors, ASICs for board-level diagnosis, computer architecture, RISC processors, cache-memory systems, and microprocessor-based systems design.

Patel received his bachelor of engineering degree from the University of Bombay, India. He received his MS in electrical engineering and PhD in electrical and computer engineering, both from the University of Texas at Austin. He is a member of the IEEE and the Computer Society.

Direct questions regarding this special issue to John Schmalzel, Division of Engineering, University of Texas at San Antonio, San Antonio, TX 78249; or e-mail at ppatel@voyager1.eng.utsa.edu.

How RISCy Is DSP?

DSP algorithms require specialized features in the processors used to implement them. However, actual DSP chips are a compromise between the resources desired and the silicon space available. This article examines the characteristics of benchmark DSP algorithms. By analyzing existing DSP and RISC chip performances, it proposes an "ideal" scalar RISC DSP chip to accommodate the algorithms.

Michael R. Smith

University of Calgary

My pursuit of an "ideal" digital signal processing chip for research led me down some unusual, but interesting, paths (see the adjacent box). The summary of that journey I present here. It is probably controversial and biased, but I hope it will spark some interest and discussion.

I describe some benchmark algorithms to establish the characteristics of DSP algorithms. I use these to suggest the features of an ideal DSP architecture, which I compare, in general terms, to current DSP and RISC architectures. Timing comparisons taken from the data books and my own research show that several on-the-market RISCs have a DSP performance close to or better than some DSP chips. My analysis of these DSP and RISC architectures leads to the suggestion for an "ideal" low-cost RISC DSP chip.

DSP benchmarks and algorithms

Setting up benchmarks for processor comparison is a fool's game. Regardless of what you choose, you will be accused (probably justifiably) of bias. The best you can do is to logically justify the benchmarks. I was familiar with the limitations connected with the chosen benchmark algorithms through my colleagues' and my research. I have considerable RISC and DSP familiarity, obtained from the university support programs of Advanced Micro Devices, Motorola, and Texas Instruments. The learning curve associated with the architecture of many sophisticated chips and their assembly language code peculiarities means that we must consider the "only so much time" principle. The DSP chips I chose have equivalent algorithms discussed (and presumably optimized) in user manuals. The RISC timings come mainly from my own research and some educated guesswork.

FIR digital filter algorithm. The finite impulse filter algorithm is representative of a number of DSP equations found in convolution, filtering, and modeling. The requirements are simple but varied. The algorithm is multiply/addition intensive and has a simple (long) loop characteristic

$$y(n) = \sum_{i=0}^{m-1} x(n-1) \times h(i); 0 \le n$$

We multiply the old and new input data values $x(n)$ by a set of m fixed coefficients $h(i)$ to form the output $y(n)$. We must fetch a group of input values from a memory array for an off-line algorithm. For on-line operation, an "infinite" amount of data must be handled, so circular buffers must be implemented. On-line operation also requires that the FIR calculation be performed quickly (between the samples) if the filter bandwidth is not to be limited. We must perform the sum operation with high precision to ensure that

continued on p. 12

0272-1732/92/1200-0010$03.00 © 1992 IEEE

Armchair designing

The topic "How RISCy is DSP?" arose from an unexpected crossover between my teaching and research interests, mixed with some armchair design and a little naivete. Some background is required as this strange combination has considerable bearing on the way I researched and present this article.

Recently, Schweber made some apparently self-evident statements about processor types.[1]

- The general-purpose complex instruction-set computer (CISC) microprocessors' rich and complex instruction set handles both basic operations and complex functions. Typically, the instructions are microcoded and take many clock cycles to complete, and their control occupies considerable silicon area.
- The reduced instruction-set computer (RISC) microprocessors' instructions are based on the assumption that the commonly executed instructions should be processed in the most efficient way possible. The result is a highly pipelined processor with the silicon used for the complex CISC control traded for additional RISC registers.
- The DSP chip performs a particular task well, and it contains all the specialized resources required to tackle that task.

These comments seemed to echo my own attempts to use modeling algorithms for communications[2] and real-time alternative magnetic resonance imaging reconstruction techniques.[3-5] After early trials with CISC chips for the hardware implementation, my colleagues and I moved through the Advanced Micro Devices microprogrammable byte-slice DSP series (extremely fast but difficult to prototype and maintain) before settling on a multiparallel board system designed around the floating-point NEC μPD77230 DSP part.[6] We also compared the capabilities of the Motorola family (integer DSP56001 and floating-point DSP96002) and the Texas Instruments family (integer TMS320C20 and floating-point TMS320C30) for the real-time modeling required for alternative magnetic resonance imaging reconstructions.

Schweber's processor classification also appeared justified in my Comparative Microprocessor Architecture course. In laboratory sessions, I attempted to illustrate the ease of using the RISC architecture for standard processing compared with the problems presented when implementing DSP algorithms. For example, many DSP applications have frequent complex memory accesses, varied instructions, and short loops (such as found in the fast Fourier transform algorithm).

We used an integer Am29000 RISC processor and an Am29027 floating-point coprocessor on the laboratory evaluation board (a STEB 29000 from Step Engineering). A major difficulty with this combination was the overhead (dead time) of sending data and instructions to and receiving them from the coprocessor, despite interesting hardware tricks such as using the address bus to transmit additional data packets.[7] This overhead, coupled with the problems of handling complex address calculation and accesses, just took away the advantage of the RISC's fast instruction cycle. The RISC appeared unsuitable for DSP, just as Schweber had predicted.

My interest in DSP applications of RISC chips would have died a natural death at this point, except that we finagled an early engineering sample of the new Am29050 floating-point RISC. This processor was totally pin compatible with the Am29000 board and avoided all the problems with the off-chip floating-point coprocessor. In addition, some specialized DSP characteristics became more apparent. For example, typical DSP chips support modulo address arithmetic, obviously not present on RISCs. However, modulo address arithmetic, circular buffers are just another way of saying virtual or physical memory translation, and the Am29050 RISC has an on-chip memory management unit (MMU) controller accessible to the compiler and the programmer.

As an armchair designer, I was naive enough to believe that—because these were the very early days of the Am29050 chip—I could get the Advanced Micro Devices designers to add additional DSP features to the next version of the chip. However, closer examination of the scalar Am29050 chip made obvious other specialized DSP architectural features, except that they were called by names different from what they are called on DSP processors. Napoleon supposedly said about his generals, "I do not want them good, I want them lucky." Did the Am29050 designers get lucky, or was this DSP capability a general property of RISC chips? My armchair examination expanded to include the scalar Sparc chip sets, the scalar Motorola MC88100, and the superscalar Intel i860 RISC processors.

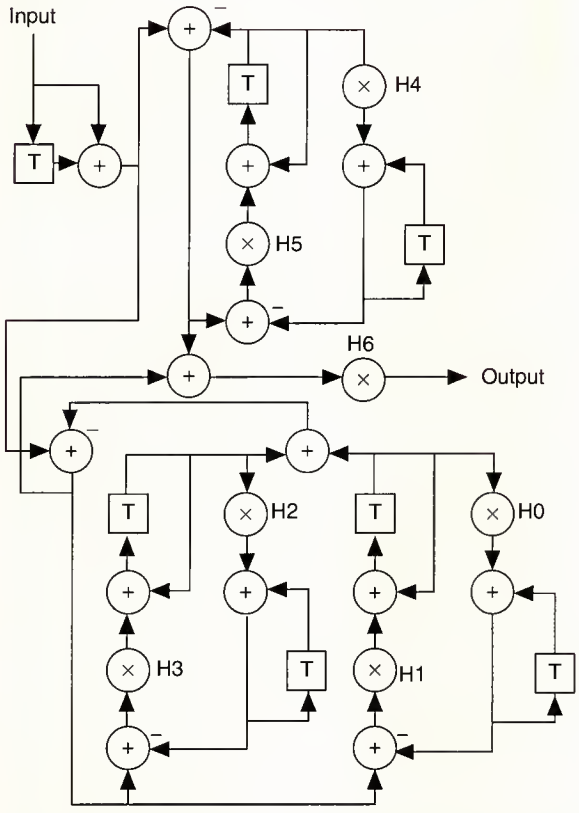


Figure 1. Schematic of a sixth-order LDI IIR filter DSP algorithm.

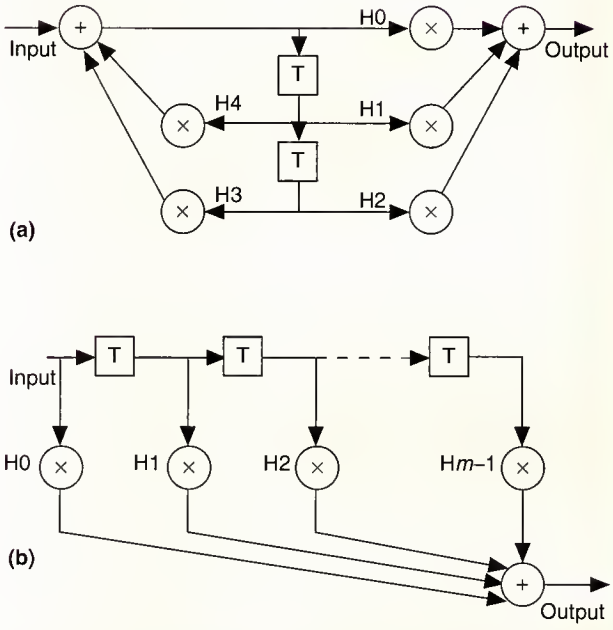


Figure 2. Schematic of the first stage of a three-stage sixth-order biquad IIR filter (a) and an *m*-tap FIR filter (b).

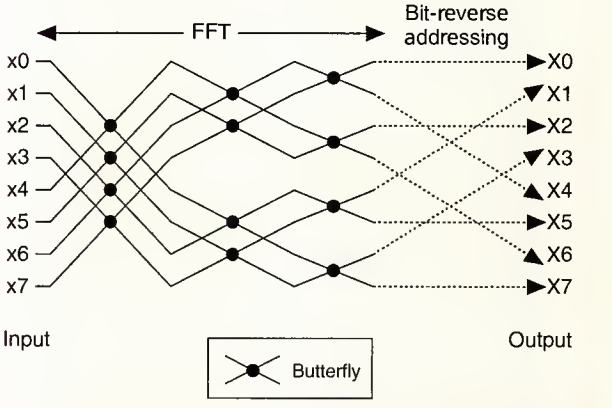


Figure 3. Schematic of a complex, radix-2, decimation-in-frequency FFT algorithm.

no accuracy is lost. It would be useful to store the fixed coefficients on chip to reduce external memory accesses.

Complex-arithmetic FIR filters have many practical applications. In addition, we can fairly easily split FIR filter algorithms into sections for multiprocessor implementation.

IIR digital filter algorithm. Continuing work on custom bit-serial Xilinx gate array technology filters[8] suggested a comparison of the effects of processor architectures on the implementation of different infinite impulse response (IIR) filter structures. Although with similar frequency characteristics, different filter structures modify the effect of quantization errors, overflow, accuracy, stability, and possible real-time speed.

Figure 1 shows the schematic of a sixth-order lossless discrete integrator (LDI) version of an IIR filter.[8] Figure 2a shows one stage of a three-stage sixth-order biquad filter, an alternate, more basic, form of IIR. These filters use a number of interrelated and order-dependent multiplication (⊗) and addition (⊕) operations. The delays $\boxed{T}$ are achieved by implementing operations of the form

$$register_T = register_{T-1}$$

at the end of each filter cycle. For comparison, Figure 2b shows the FIR filter schematic using the same notation.

The IIR algorithms are good benchmarks as they are characteristic of the class of DSP algorithms involving simple short

Requirements of the "perfect" DSP architecture

- Fast instruction cycle—this is different from high clock speed
- Cycle time adjustable according to instruction type
- Fast hardware multiplier
- Floating point for easier algorithm design
- High precision, implying wide data buses for memory, internal processor transfers, registers, and on-board processing units
- Several data buses available to reduce memory bus conflict/transfer overhead
- Harvard architecture and/or instruction cache to avoid instruction and data-fetch clashes
- Duplicate resources for parallel computation of real and imaginary components of complex numbers
- Zero branching overhead
- Dedicated hardware required for address calculations to avoid APU resource clash with the main algorithm
- Extensive temporary registers needed to reduce unnecessary fetches of continually used data
- Fast and reliable, easily programmed, developed, and upgraded
- Inexpensive and easy-to-develop peripherals
- High level of customer support
- Inexpensive to purchase
- Lower power consumption with a standby mode

loops where loop overhead can be a problem. An advantage of a RISC is its highly pipelined nature. If the various on-board pipelines can't be kept filled because of data dependencies, this becomes a disadvantage for short loops or short program sections. Also, the IIR DSP algorithms continually reuse a considerable number of temporary variables, so loss of precision and fast access become important considerations.

FFT algorithm. Figure 3 shows the schematic for a complex number, radix-2, decimation-in-frequency fast Fourier transform (FFT) algorithm. (The literature provides a wealth of information on the FFT algorithm, but for a tutorial see Burrus and Parks' book.[9]) The basic FFT element is the butterfly

$$A'(m) = A(m) + B(n)\ W(p)$$
$$B'(n) = A(m) - B(n)\ W(p)$$

Typically, large (1,024) arrays of complex variables (A and B) and fixed coefficients (W) are involved. The address calculations are not straightforward, and the memory accesses are numerous. As part of the FFT passes, the data positions must be reordered. For example, in a 256-point, radix-2 FFT, array location 203 (%11001001) must be moved to location 147 (%10010011)—bit-reverse addressing.

This algorithm is representative of the "complicated" class of DSP algorithms. The overall DSP characteristics of the processor are very systematically tested. The many multiplications and additions are linked, but they are not of the simple multiply-and-accumulate format found in the FIR filter. There are also a number of loops, including some tight inner loops. The number of registers (or data cache) available for use as address pointers, constants, and variables becomes important. Both integer and floating-point operations are needed. (As a moot point, when implemented on a RISC processor, do these FFT algorithms become fRISCy Fourier transforms?)

Real chip architecture compromises

The Requirements box identifies what I consider to be the desirable DSP features of a processor suitable for handling the benchmarks. Custom design is probably the only way to obtain everything. Microprogrammable byte-slice DSP products[6] are fast, but by no stretch of the imagination easily programmed or upgraded. With today's technology we might achieve a fast, reliable custom design using available library modules, but at a high development cost. For these reasons it is better to make some reasonable compromises on the "perfect" DSP system and examine currently available RISC and DSP processors. Since these chips are compromises, different processors may give maximum performance for different applications, leading to apparent biases in the choice of DSP benchmarks.

Fast instruction cycle. Dedicated DSP chips typically have an instruction time twice as long as the clock cycle. In this time the chips perform many parallel operations, including memory access(es) and calculation(s). By comparison, the ideal RISC would initiate and complete a simple instruction(s) every clock cycle. This is *not* the same as saying that each RISC instruction takes only one clock cycle.

For example, RISC and DSP FADD (floating-point add) and FMULT (floating-point multiplication) instructions take between 75 and 150 ns to complete (the equivalent of three to six clock cycles). They require a heavily pipelined arithmetic processing unit (APU) for efficient operation. A new floating-point instruction can be started and completed every instruction cycle, provided the pipeline can be kept full, giving the RISC an advantage with its faster instruction cycle. Typically, DSP algorithms are very repetitive (for example, FIR filters) or have many things to calculate (IIR filters). Thus, it normally is not a problem to keep the APU pipeline full. However, to avoid possible delays in the RISC memory pipe-

In some ways DSP chips resemble a CISC processor in which the instruction with the longest execution time controls the cycle time.

line, the chip must store intermediate results rather than send the temporary values out to slower external memory.

Efficient use of instruction time. The DSP chip's instruction cycle is very busy. The chip needs time to calculate addresses and fetch data from on-board RAM, in addition to the time to actually carry out the calculations. In some ways current DSP chips are CISCs in which the instruction time is controlled by the instruction with the longest execution time. On the other hand, the RISC is pipelined everywhere to minimize its instruction cycle. The RISC has the advantage here as the instructions are faster and a number of high-speed instructions can be brought together, *if and when needed*, to perform complex DSP instructions. But a RISC loses its advantage if we have to compound too many instructions to emulate a complex DSP instruction.

Loop overhead. The branch instruction requires fast and decisive action because this control operation is a waste of processing time for any chip. The typical solution in DSP architectures is dedicated hardware that allows zero-overhead loop(s). However, the compromise is that there is often only enough silicon space for the hardware to handle the inner (possibly only) loop.

How do RISCs handle this problem? In typical computer fashion, designers trade off speed and memory usage. First, the chip has special hardware to handle loops in general. RISC branches cause a break in the instruction pipeline as a nonsequential fetch is performed. We can overcome this loss of time with the delayed jump instruction (always processing the instruction(s) after the jump). On-board branch target caches store the following instructions, keep the instruction pipeline full, and produce low-overhead loops (another example of the specialized DSP architectural features present in RISCs but under other names).

However, this leaves the RISCs with one or more additional branch test instructions every time around the loop, and DSP algorithms loop often. RISC cycle time is faster than DSP cycle time, and for a single-cycle test, decrement, and jump instruction, very little loop overhead occurs. The exception is the single instruction loop found in DSP algorithms such as the FIR filter. However, we can trade the loop overhead for increased program length by using straight-line coding or, more efficiently, by grouping a number of the loops. For most programs, we should expect an increased RISC code length compared with that of CISC and DSP processors.

Multiple data and memory access buses. The RISC and DSP chips we investigated contain roughly equivalent multiple data and memory access buses to allow free movement of data, but their configuration differs. The chips resolve the conflict of data and instruction fetches in a variety of ways. Methods include large data caches (to free the data bus for instruction fetches), separate instruction and data buses, multiple-data buses, branch target caches, large register banks, multiported memory, and register preforwarding. However, in a private communication, Motorola engineer Wei Chen brought up a point to consider: the effects on high-speed performance of using single-cycle external (burst) memory and a large register bank (Advanced Micro Devices Am29050 RISC), dual-ported on-chip memory (Motorola DSP96002 DSP), or an on-chip data cache (Intel i860 RISC). Use of a cache means possible nondeterministic hit rates with implications in the design of real-time systems.

Multiplication-extensive algorithms. The floating-point RISC and DSP chips all have an on-board hardware multiplier (typically pipelined). RISCs have a multiplication time similar to DSP chips but can complete pipelined multiplications at a higher rate because of their faster instruction cycle. Both processor types have additional resources that allow integer, floating-point add, and floating-point multiplications to complete in parallel for faster operation. The repetitive nature of the DSP algorithms lessens any pipelining problem, but code ordering (via an optimizing DSP-specific compiler) and correct use of the floating-point registers can make a considerable difference in speed. The availability of register forwarding to shortcut the pipeline is important. The RISC sets with an off-chip multiplier coprocessor can have considerable overhead from accessing the coprocessor.

However, the presence of the multiplier is not the only important factor. The chip must be able to quickly move data in and products out of that multiplier. Many integer DSP chips have one destination for the multiplier result, which is a severe programming bottleneck. In addition, the sources of the data are often severely limited. This is not so much a problem with the current generation of floating-point DSP chips. It is, however, very easy to choose an algorithm in which the number of floating-point sources and destinations are insufficient (because of the need to store intermediate results). For example, the Cypress and LSI Logic Sparc chips and Motorola MC88100 have only 30 registers attached to the multiplier (compared with 192 on the Am29050 processor). These registers must be reloaded continually from slower external memory, requiring unproductive instruction cycles. The i860 also has only 30 registers. However, it has a limited (but normally sufficient) ability to switch into a dual-instruction

mode. This mode lets it use one register group while another reloads—through a (limited) simultaneous quadruple-register load mode—from a reasonable-size on-chip data cache.

Many RISCs do not have the multiply-and-accumulate (MAC) instruction present in most DSP processors. This very important instruction uses the multiply/add pipeline to let a single-instruction fetch efficiently initiate two simultaneous floating-point operations. We can implement it using separate FADD and FMULT instructions, but often with a significant loss of speed. However, some RISCs' APUs are constructed so that for short DSP loops the filling and flushing operations of the MAC pipeline are time consuming. Other RISC processors lose efficiency because they do not permit a wide range of MAC instructions (compare the different MAC variants necessary in the FIR and FFT algorithms).

Precision. RISC and DSP architectures use various methods to offset precision problems. In evaluating the precision of DSP and RISC architectures, we took the following factors into account: the memory bus data width, the internal register data width, the internal bus data width, and the data width of the on-chip processors.

Even APU design must be considered. For example, we might expect problems on the DSP56001 with its 24-bit-wide on-chip memory. After multiplication, a result would be 48 bits wide, truncated back to 24 bits when stored to external memory. This would lead to serious error propagation during subsequent passes through an FIR algorithm. In fact, the DSP56001 avoids these problems by using 56 bits for sums-of-products operations and saturation arithmetic on storage. By contrast, the Am29050 RISC processor gets high-precision capability through dual (64-bit) register access capability to the APU.

A problem with integer processors is that they must scale the data when there is a possibility of overflow (for example, during the FFT butterfly calculations). The overhead for overflow checking is considerable, so the standard approach is to scale automatically, yielding a result whose accuracy is normally not optimum. In a Motorola FFT application note, Sohie suggests that the optimal scaling of all arithmetic results is obtained by implementing floating-point DSP algorithms.[10] Floating-point operations use a different scaling factor for every number, so scaling and loss of accuracy only occur as and when necessary. In the floating-point processors, this scaling occurs without additional time penalty. (The word "additional" is important because the instruction cycle time for the floating-point processors can be slower than for the integer processors.) In addition, floating-point DSP algorithms are frequently easier to design and implement than integer ones.

Complex arithmetic. None of the RISC or DSP chips we surveyed had dedicated complex-arithmetic resources. The silicon overhead is too expensive for the duplication of normally unused resources. The existing architectures meet most requirements for complex arithmetic, except that we can expect two to four times slower performance than with the equivalent real-arithmetic algorithm. The ability to bring a pair of memory locations (for example, sine and cosine values) into adjacent internal registers in parallel with other operations, as in the i860, would be useful on other RISCs.

Variants of pipelined multiply-and-add instructions are important for efficient RISC implementation of DSP algorithms.

Duplication of resources such as ALUs and multipliers is really a parallel processing aspect of DSP algorithms. The DSP96002, MC88100, and i860 are designed for multiprocessor operation, and multiple Am29050 processor board designs have been reported. Texas Instruments has introduced the TMS32040, which has handshaking capability with six other processors. Efficient complex-arithmetic algorithms implemented on a multiple-processor system would need dual-ported memory. This is necessary to implement the crossover data paths to give the real and complex processors access to the same variables. (I don't have the experience to compare the success or relative ease of programming RISC and DSP chips for complex-arithmetic multiprocessor applications.)

Standard address calculations. Address calculations are a possible serious time consumer on a RISC and come in a number of forms. Straight-line coding is better than indirect register instructions for addressing multiple internal registers on both RISC and DSP chips. RISCs obviously are deficient in the use of incrementing addressing modes for the DSP algorithm, but they do not perform as poorly as we might expect. The AMD RISCs have LOADM (load multiple instruction), so the on-chip MMU handles the address autoincrementing and efficiently uses burst-memory capability. The Sparc chip set and MC88100 have an offset addressing mode that allows the (straight-line) coding of the autoincrementing mode, taking advantage of the faster RISC instruction cycle. By contrast, the superscalar i860 has such an extensive addressing mode capability that a RISC purist would probably consider it obscene.

Frequently, we can avoid address calculation by using a direct memory access unit for block data moves, often in parallel with other CPU operations. DSP chips often include this feature. Frequently, the manuals are unclear about what conflicts (transparent stalls) occur when the CPU accesses a

Basic RISC programming considerations

- Each RISC instruction is pipelined—fetch, decode, execute, write back.
- Integer operations have a single-cycle execute phase.
- Floating-point operations have the execute phase pipelined over many cycles, but one can be initiated and completed every cycle.
- Instruction and data memory access are pipelined, even from the cache.
- The integer and floating-point operation(s) are independent.
- Floating-point multiplication and addition units can operate independently.

block of on-chip data memory at the same time that the DMA unit tries to update the memory. RISCs can typically relinquish the data bus to allow an off-board DMA controller to move data between external memory blocks. As RISC chips can have an on-chip instruction cache or a separate instruction bus, the external DMA moves can occur in parallel with other on-chip operations. DMA might be a useful approach for handling specialized RISC addressing requirements such as bit-reverse addressing.

When we cannot avoid RISC address generation by grouping the memory accesses, we must simply calculate the address at the expense of additional cycles. The overhead is frequently only half what we might expect: The same address calculation is often used both for loading and storing values—for example, with an in-place FFT algorithm. A large number of on-board (integer) registers on a RISC allows the storage of these addresses for later use.

Specialized address modes. The implementation of the circular buffer operation (real-time FIR applications) is a special case of address calculation. The DSP chip's buffer is typically implemented using modulo arithmetic in specialized hardware to allow parallel operation with no resource conflict with other operations. RISCs, by comparison, would require considerable software overhead to calculate new addresses and their adjustments within the bounds of the circular buffer. However, RISCs typically have virtual memory management capability, either on chip or as part of the chip set. When this capability is not automatic, we can map several virtual memory blocks into a single physical memory space. This provides a low-overhead circular buffer operation, although without the same flexibility as the modulo arithmetic address capability. For small circular buffers, the on-chip register window available on some RISCs can be put to good use, provided the window-handling instructions are flexible enough, and the register window has direct access to the on-chip multiplier (compare the Am29050 and Sparc chips).

Other specialized addressing modes (for example, the bit-reverse addressing used in the FFT algorithm) are standard for DSP chips, often with zero overhead. RISCs have nothing similar. RISC bit-reverse addressing can be handled by fetching the addresses from external memory and then mapping the address information into a physical space using the MMU. This approach still requires considerable overhead.

Storage of intermediate results. The floating-point RISCs and CISCs store intermediate results in different ways. DSP chips have fewer registers but more on-chip data memory than do RISCs. Some RISCs have 30 to 200 registers that often can be used as a small memory block when the register values can move directly into the multiplier or APU at high speed. Other RISCs have an on-board data cache. The integer DSP chips we investigated do not have enough destinations for APU results, so storage of the intermediate results incurs considerable overhead. RISCs also need many registers for temporary storage of addresses to overcome their lack of address mode capability.

Ease of use. The RISC and DSP chips we investigated are fast and reliable, and commonly have on-board timers to generate interrupts for real-time DSP. DSP chips have the advantage of on-chip serial ports and standby power modes, whereas these are additional external hardware for RISCs. However, adding such items would be fairly straightforward. The floating-point processors tend to be more power hungry than their integer counterparts. Again, this is probably more current usage (a pun as well as a problem), and we can expect a trend toward lower consumption on all chips with the introduction of 3V systems.

Users need application notes that show concepts and get development going quickly. DSP chip manufacturers provide the application notes on DSP algorithms, though the RISC manufacturers are still new at this game. (This means I can make a fortune writing DSP application notes for RISCs, so again RISC comes out ahead.)

The actual instruction set and its ease of use also helps users. Unlike the designers of the easier-to-understand Am29050 instruction set, the i860 code designers appear to have worked on the old principle, "If it was difficult to design, it ought to be difficult to understand." However, what mnemonics would you use to distinguish between the numerous MAC instruction variants found on the i860? When efficient DSP compilers become more available for RISCs, this will be less relevant.

Texas Instruments DSP chips are downward compatible over a wide range of products. This approach provides the advantages of a load of happy previous users and a big software base, but it puts a huge burden around your neck if you are trying to add a new feature to an existing chip. On the other hand, I believe RISCs are still in their infancy so that manufacturers can include new features without alienating existing users.

Efficient RISC programming for DSP applications

- Split the DSP algorithm into many interleaved tasks or data streams to keep pipelines full.
- Choose the algorithm form that is best at combining the data streams and keeping a full pipeline.
- Move integer operations (for example, address incrementation) into the (transparent) stalls associated with floating-pipeline flushes.
- For short DSP loops, fold two adjacent loops to fill one pipeline as another flushes.
- Use efficient single-cycle loop instructions via "down counting."
- For really short loops, use straight-line code.
- Avoid data dependencies or excess external memory access using (many) registers to store intermediate results.
- Look out for the weak scalar RISC feature—reloading the registers from external memory.
- Bring in data blocks from "burst" memory to avoid both memory and APU pipeline problems.
- Use the MMU to avoid address calculation, for example, in circular buffer addressing.
- Invest in a good, intelligent DSP RISC compiler.

RISC and DSP architectures for DSP

I've examined the characteristics of DSP algorithms in general terms—that is, in terms of the required architectural features of a processor. It is obvious that RISC processors have some DSP capability, but can they provide the performance? Does the question, How RISCy is DSP?, apply to the current RISCs or do some additional features need to be added? The time has come to pay the piper.

The basic programming of a RISC processor is fairly straightforward because the instructions are "simple." However, one reason DSP processors are so efficient for DSP algorithms is that the code is adapted to make the best use of the processor's architecture. Intelligent use of the RISC architecture will lead to similar efficiencies. The Basic RISC box details the architectural features we must consider when programming a RISC processor. The techniques for efficient RISC programming are fairly straightforward, once spotted. The Efficient RISC Programming box details the techniques used in this article, and they can be compared with the approaches used in developing optimizing compilers. (Booth provides a tutorial review[11,12] on optimizing compiler techniques in scalar and superscalar RISC contexts.)

Benchmark reference points. The timings for the benchmarks—based on user manuals and application notes from Texas Instruments,[13-15] Motorola,[10,16] and Intel[17,18]—are probably accurate. A number of results were scaled according to the last available processor clock speed. For better comparison with the RISC figures, I have changed some DSP processor examples to straight-line coding from a "looped" form when this would give an improved performance figure.

Manufacturers' comparisons pose two problems. First, there is running under special environment (RUSE), where it appears that two timings are identical, but in fact they are not. For example, one manufacturer may base a set of timings (say, for the FFT) on the data already in the cache or on-chip memory, whereas another manufacturer bases timings on data starting in external memory. The timing adjustment for this is difficult because many DSP processors (and some RISCs) have load/store operations that work in parallel with other operations, and perhaps the importance of data placement depends on the application. High-speed memory is expensive, and the timings will change with the memory configuration (compare N wait states with burst memory).

A second problem is running out of resources as the number of points handled by the DSP algorithm increases. The TMS320C25 manual's FFT timings offer a clear case of the effect of lack of resources.[19] The on-chip memory can handle the 256-point, complex FFT (1.8 ms). Scaling the time to 1,024 points should give 9.0 ms, but the processor in fact takes 15.6 ms as the data must now be fetched from external memory. Breakpoints occur after the 95th FIR tap on the Am29050 RISC,[20] after 512 points in the radix-4 TMS320C30 FFT,[19] and after 1,024 points in the radix-2 i860 FFT algorithm. Many timings in the literature are taken when the processor is just at the edge of running out of resources, and a drastic loss of performance would result if the DSP algorithm used additional points. (It would be very convenient for the reader if the application notes would *clearly* point out where the breakpoints are, rather than gloss over them.)

As the DSP use of RISC architectures is not de rigueur, I made theoretical calculations of the expected times and then experimentally checked them where possible. The i860 timings are based mainly on Intel's publications.[17,18] I timed the Am29050 processor on an 8-MHz Step Engineering STEB evaluation board with overlapped instruction and data buses, and zero-wait-state memory. The overlapped buses do not take full advantage of the Am29050 processor architecture when extensive memory access occurs. However, the STEB board nicely simulated a Sparc system, which has overlapped instruction and data buses. The inaccuracies associated with this overlap are important only when there are a large number of memory fetches (for example, for the FFT algorithm

Table 1. Timings for DSP algorithms on various processors.

	DSP				RISC				
Description	TMS320C25	TMS320C30	DSP56000/1	DSP96002	i860	MC88100	Sparc	Am29050	Ideal
Type	Integer	FP	Integer	FP	FP	FP	FP	FP	FP
Clock speed (MHz)	50	40	33	40	40	30z	25	40	50
Instrument cycle (ns)	80	50	60	50	25	30	50	25	20
FIR filter									
Cycles	$N+8$	$N+5$	$N+7$	$N+7$	$1.12N+30$	$(7N+20)/2$	$(7N+20)/2$	$N+15$	$N+15$
91 tap time (μs)	7.92	4.0	5.88	5.0	3.3	9.86	16.43	2.65	2.12
Cycles	$N+8$	$N+5$	$N+7$	$N+7$	$1.12N+30$	$(7N+20)/2$	$(7N+20)/2$	$2N+28$	$N+28$
191 tap time (μs)	16.56	8.2	11.88	10.0	6.1	20.35	33.93	10.25	4.38
IIR filter									
LDI cycles	--	25	20	22	29	26	26	26	20
LDI time (μs)	--	1.25	1.0	1.1	0.73	0.78	1.3	0.65	0.40
3 biquad cycles	$15N+4$	$24+6N$	$5N+1$	$5N+5$	31	43	26	28	23
3 biquad time (μs)	3.72	1.8	0.8	1.0	0.78	1.29	2.05	0.70	0.29
Radix-2 FFT									
256, complex (ms)	1.8	0.68	0.94	--	0.18	0.98	1.38	0.63	0.36
256, bit reversed	--	--	--		0.20	1.1	--	0.79	0.36
1,024, complex (ms)	15.6	1.97	4.72	1.04	0.97	4.70	6.66	3.08	1.73
1,024, bit reversed	--	--	--	--	1.11	5.19	--	3.47	1.73
Radix-4 FFT									
256, complex (ms)	1.2	0.53	--	--	--	--	--	0.44	0.26
256, bit reversed	--	--	--	--	--	--	--	0.54	0.26
1,024, complex (ms)	--	2.53	--	1.81	--	--	--	2.13	1.2
1,024, bit reversed	--	--	--	--	--	--	--	2.52	1.24

with the Am29050 processor). The Sparc floating-point pipeline is not detailed in the Sparc definition.[21] I assumed a pipeline equivalent to that on the other RISCs, but this assumption makes the Sparc timings suspect.

The DSP algorithms are basically collections of integer address calculations, memory fetches and stores, and floating-point operations. With RISCs having very similar floating-point instructions, except for the MAC instruction, many of their timings are validly based on the experimentally verified Am29050 processor timings. Provided we can keep the RISC pipelines full and use single-cycle memory, there are no differences in the number of cycles required, although the coding order may be very different. Variation in required cycles occurs only when additional external memory accesses are necessary because the particular RISC has insufficient registers to store intermediate and reused values. Because of my lack of familiarity with some of the RISC processors, timings differing by less than 10 percent are probably equivalent.

Basic RISC differences

Table 1 gives the timings for the various algorithms. The variations in timings on the RISC processors occur because

of a number of fundamental architectural differences.

The MC88100 has only 30 registers for floating-point and integer variables compared with over 100 registers for other RISCs. The lack of registers often translates into additional (nonproductive) memory fetches and stores as temporary values have to be moved. The floating-point Am29050 processor has a large register bank attached to the floating-point multiplier, which helps increase speed. This bank acts as on-chip memory for filter coefficients and state variables, avoiding memory access overhead except when the number of coefficients is very large. The i860 dual-instruction capability permits the loading of one register group while another is used, overcoming the problem of its limited number (30) of floating-point registers attached to the APU.

The register-plus-offset addressing modes of the MC88100, Sparc, and i860 chips offer a minor advantage over the load multiple LOADM Am29050 addressing mode in some algorithms.

When we could organize the algorithms into sections in which three or more products were summed, the Am29050 and i860 processors, with their MAC instructions, had a distinct advantage. However, effective use of the MAC instruction in short loops was often next to impossible because of this instruction's deep pipeline. The i860 APU seemed more inefficiently designed than that of the Am29050, particularly when we attempted to clear and set up the MAC pipeline. However, the wide variety of MAC instructions available with the i860 often gave it back the advantage.

Getting "good" floating-point performance was fairly straightforward on the Am29050 processor because its floating-point instructions are implicitly pipelined. By comparison, the i860 had explicit slow scalar (program and forget) and fast vector (pipelined) floating-point operations. "Best" floating-point performance on both the Am29050 and i860 processors required detailed knowledge of the APU architecture (hence the need for intelligent DSP compilers).

The varying floating-point pipeline depths of the various chips had a drastic effect on the way we coded the algorithms. RISC processors with long pipelines and few temporary registers (for example, the MC88100) were penalized by the pipeline problem and were more difficult to program to avoid the (transparent) stalls, unless highly repetitive actions (long loops) were available.

We coded all RISC loops so that the flushing of the pth loop pipeline intermingled with the filling of the $p+1$th loop pipeline. Tutorial articles provide more detailed information on the advantages and limitations of implementing various DSP algorithms on RISC processors.[20,22,23]

FIR filter comparison. An N-tap FIR filter is based on a single-instruction loop requiring typically three or four instruction locations on DSP chips. By comparison, RISCs had to be straight-line coded, requiring between N (Am29050 and i860 processors) and $3.5N$ (Sparc and MC88100 processors) instruction locations on RISCs. The difference in RISC program length depended on whether or not the data and filter coefficients could be stored on chip or had to be continually fetched from external memory. If the deep RISC APU pipeline is not properly handled, each instruction may take three or four cycles, again demonstrating the need for an intelligent DSP compiler.

RISC results show the advantages of the MAC instruction and easy access of registers to the multiplier (as on the Am29050 and i860 processors). Using a RISC register window as a small on-chip circular buffer or dual-instruction capability offered a considerable advantage over bringing data from an off-chip buffer implemented with the MMU. This was true only if the register window had access to the multiplier (compare the Am29050 and Sparc processors). The main difference between the top-performing RISCs was that the i860 lacked the Am29050 processor's MAC variant A×B+0 instruction. This necessitated a less-efficient FPU pipeline start-up (and also flushing). We assumed the application involved a floating-point FIR using integer A/D values, which penalized the i860 with no explicit integer-to-float conversion instruction. The scalar Am29050 processor outperforms the superscalar i860 for a small number of taps, but the i860's dual-instruction capability comes into its own for large tap numbers, which require access to more filter coefficients.

IIR filter comparison. The sixth-order LDI filter is just before or after an insufficient resources breakpoint for many RISC and DSP chips. Because of the TMS320C30's restrictive addressing requirements on parallel operations and low number of floating-point registers, a higher order LDI filter would require additional nonparallel memory stores.[24] The DSP96002's parallel operations are less restrictive,[16] and its breakpoint would occur much later.

Adjusting the Am29050 processor register window to implement delays and using the MAC instruction offered no advantage for the LDI structure and a minor (three-cycle) advantage for the biquad structure. More important on the biquad structure algorithm was the ability to store multiplier coefficients and state variables on chip. The RISC requirements of additional temporary registers for intermediate results and the overlapping of loops to avoid pipeline stalls had to be considered. The MC88100 and Sparc required cycles equivalent to those required by the Am29050 and i860 for the LDI filter, but simply ran out of registers for the biquad filter. They required reloads from external memory (10 and eight on the MC88100 and Sparc processors). Clearly, the algorithm implementation will have a significant bearing on the success of a RISC in a DSP application.

The integer Am29000 RISC/Am29027 floating-point coprocessor combination performed poorly because the coprocessor instructions must be stored in an Am29000 processor register for fast transmission. This makes fewer registers available for temporary variables. The large number of different

Overlapping multiple memory fetches with floating-point operations, rather than dual instructions, gives superscalar RISCs a major advantage.

floating-point instructions required for an IIR algorithm indicates that an intelligent coprocessor capable of directly reading the instruction stream (for example, the Sparc coprocessor) or an on-board multiplier is required.

FFT analysis. The timings should be taken with a grain of salt as they are not all really equivalent (the RUSE problem discussed earlier). For example, the Am29050 processor timings are based on the data starting and ending in external memory (because they *have* to be there). Other results may assume that the data are already on chip. Whether this is fair or not depends on your application. FFTs are often just part of a DSP program. If the stage after the FFT needs the data on chip, why include the time for moving off chip and then back on?

We could improve the FFT timings of RISCs considerably by avoiding all address calculations and straight-line coding the algorithm. While this might be a practical solution (as in the DSP TMS320C25 processor code), it avoided exposing the limitations of RISCs in DSP applications. Therefore, we implemented a looped algorithm with address calculations. Even so, the program for RISCs was considerably larger than for DSPs. (For example, an FFT butterfly was performed in four instructions on the DSP96002 and 10 instructions on the TMS320C25, compared with 18 instructions on the Am29050 RISC processor.)

Overlapping a number of FFT butterflies was de rigueur for any high-speed performance on the processors. For example, we overlapped four for the Am29050 processor. The best FFT timings from the literature include some optimized passes. Both radix-2 and radix-4 timings are given. We assumed that the sine and cosine coefficients were precalculated (otherwise, double the calculation time).

The use in the FFT of the DSP96002's single-cycle simultaneous FADD/FSUB instruction offered such a peculiar advantage over the two to three cycles for the equivalent operation on the other RISC and DSP chips that I predict this instruction will become more common (dirty pool, Motorola).

The timings clearly show that the lack of specialized addressing modes (bit-reverse addressing) has a considerable effect on a RISC's DSP performance.

The i860's dual-instruction capability gave it some advantage over the scalar Am29050 processor. However, the i860's ability to overlap multiple memory fetches with floating-point operations and the provision of many MAC variations were its main advantages.

The very deep FADD (5 deep) and FMUL (6 deep) pipeline of the MC88100 caused some difficulty in the efficient coding of the inner butterfly loop. Pipeline depth is not defined in a Sparc chip's architectural standard, so a DSP program efficient on one Sparc set may be inefficient on another. Also, not all Sparc sets have an on-board instruction cache, so there may be considerable conflict as data and instructions compete for a single data bus.

Again, the MC88100 did not have enough storage to keep all floating-point variables and address pointers on chip with only 30 registers. It needed additional store/load memory cycles. The Sparc and Am29050 had ample address pointer space using their register windows. The i860 avoided the difficulty via its dual-instruction capability and integer register bank. Here is a specific illustration of the problem: C-compatible, efficient radix-2 and radix-4 implementations on the Am29050 processor used 52 and 130 of its 192 registers.[23]

CRISP—A future ideal RISC DSP chip?

I now introduce my concept of an "ideal" RISC DSP chip: the comprehensive reduced instruction-set processor—Smith's CRISP. (The acronym suggests a processor that is neat or hot, that is, fast. UK readers should appreciate the double word play. For the less world-traveled, "crisp" is the English word for potato "chip" and, unless the recession has really struck hard, Smith's is a major "chip" manufacturer.)

A fundamental problem with using a RISC as a DSP processor is cost. Prices for cost-reduced DSPs are around $50, while current DSP-suitable RISC processors can be four or more times that price. To achieve economy of scale, a large number of CRISPs must be sold. This implies usage in both DSP and other general applications, with different clock speed processors available. Superscalar (dual-instruction) architecture is overkill for most general situations, so the CRISP will have a basic scalar RISC core with the best features from existing RISCs. The benchmarks showed that all RISCs were similar, provided they did not run out of resources, especially temporary registers. The presence of many MAC instruction variants gives a distinct advantage.

Basically, I imagine the CRISP as an Am29050 processor with its 192 (partially windowed) registers and implicit floating-point pipeline instructions, supplemented with features stolen from the i860. This combination would address the major failings of the current RISC architectures for DSP applications. If any RISC chip set designer takes DSP to heart, these features would be part of that chip's repertoire. AMD has recently gone some of the way toward putting additional features helpful for DSP into RISC processors with the intro-

duction of the integer Am29200 microcontroller. This microcontroller has a basic Am29000 RISC core with serial, parallel, and printer "video" ports, together with control logic to permit easy handling of peripherals (including some DMA action). Add a low power consumption (standby) mode and an on-board integer multiplier to the current Am29200 processor or provide a Am29050 processor upgrade, and you will be well on the way to a CRISP. However, as the benchmarks showed, the CRISP is a bit more than that.

The CRISP ALU architecture would allow a wide variety of MAC instructions for full use of the FMULT and FADD operations implemented by different resources that can operate in parallel. It was this feature, rather than its dual-instruction capability, that often gave the i860 a performance advantage over the Am29050 processor. The CRISP ALU would also be versatile in the way it collects partial sums generated from various data streams. This was an area where both the i860 and Am29050 processors lost performance for short DSP data sections and loops. However, we should hire the Am29050 instruction designers to develop the mnemonics rather than the designers from Intel. (As I said before, the very versatility of the i860 APU is part of the problem.)

The i860's ability to handle simultaneous memory access in conjunction with arithmetic operations rather than actually issuing the dual instructions simultaneously often gave it an advantage. Therefore, we should add some new load/store instructions to the CRISP to be used in parallel with the ALU operations. Naively examining the data flow diagram and APU diagrams in the Am29050 user manual[25] gives the impression that suitable data paths are already present in the Am29050 processor and could be easily activated. (According to P. Eichenseer of AMD's 29K-hotline, the actual activation difficulties are associated with data dependency control and required duplication of certain resources.)

The CRISP DMA_LOADM instruction loads a group of N registers from external memory while allowing simultaneous (floating-point) operations on a different bank of registers. Using this instruction with an intelligent compiler or assembler would allow N floating-point operations and N memory loads to occur in $N + 2$ clock cycles—a time saving of 50 percent for DSP algorithms that extensively use memory access. Even as low as 8, N would be large enough to gain a considerable advantage.

A useful, but not necessary, single-instruction LOAD2 variant could load double-precision operands. With most RISCs, this addition would complement their implementation of double-precision arithmetic operations. This instruction would also help complex-arithmetic memory accesses. The equivalent instruction already exists on the Sparc and MC88100 with their offset addressing mode but, rather than being a single pipelined instruction, requires multiple instruction fetches.

Timing showed that a 1,024-point, complex FFT algorithm uses considerable time for bit-reversed address manipulation. Although we could modify the MMU, specialized DSP requirements might be best addressed (another pun) by an external DMA controller or other logic. This would keep the CRISP chip price lower for more general applications. The existing Am29050 processor LOAD instructions already have the control operation to access a number of addressing spaces and might be used to activate these external devices. To allow manipulation of the on-board MMU hardware to implement such items as circular buffer operations, the CRISP would have the ability to turn off automatically handled MMU operations.

We attempted to simulate CRISP performance. The last column of Table 1 gives the results. The similarity between the CRISP and the Am29050 processor let us generate "experimental" timings. Using the new CRISP DMA_LOADM to load floating-point registers with simultaneous floating-point register use requires overlapping loops in the DSP algorithms, a technique already discussed to minimize the effect of pipeline dependencies. Therefore, we could simulate the DMA_LOADM instruction by setting up the Am29050 processor LOADM instruction but replacing the actual LOADM instruction with a NO-OP instruction to remove memory access dependencies.

The new MAC instruction variants would play a significant role only in the FFT algorithm and have a lesser effect on the IIR filter timings. They would require overlaying the FADD instruction of one loop with the FMULT instruction of another loop. (Again, this is only a minor modification to the program, as the loops are already overlapped to reduce pipeline dependencies.) We simulated this overlaying by removing the FADD instructions from the appropriate code section. We simulated improvements associated with improved specialized addressing simply by deleting the corresponding code sections. I therefore have fair confidence in the CRISP timings. The timings are probably an overestimate because the effect of the overlapped instruction and data buses on the STEB evaluation boards was not removed. For the memory-intensive FFT algorithm, the effect on the timing is considerable (I estimate 20 percent).

THE CURRENT SCALAR AM29050 and superscalar i860 RISCs consistently outperform the other RISCs investigated. This is because they have a large register window bank (or can implement the equivalent via a dual-instruction mode) connected to the floating-point processor unit and have a multiply-and-accumulate instruction. Manufacturers need to reduce the RISC's power consumption and cost if they are to grab a niche in the telecommunications market. Specialized RISC compilers for DSP applications will be needed to optimize loop, register, and pipeline operations. A combination of the best of the i860 and Am29050 RISCs would make a very practical scalar comprehensive reduced instruction-set pro-

cessor (CRISP) DSP chip, especially if the features of the new Am29200 microcontroller were added.

In a recent article, Wilson[26] suggests that "just as RISC chips have displaced CISC in general applications, so perhaps will DSP processors." However, perhaps it is RISCs that will displace DSP processors: I contend that some existing RISCs are already more than centrally placed with respect to the performance of the current dedicated DSP chips. Therefore, it is worthwhile to "take a RISC in DSP" or perhaps even to "(number) crunch on a CRISP."

Acknowledgments

The University of Calgary, Canada, and the Natural Sciences and Engineering Research Council (NSERC) of Canada supported this work. University support programs from Advanced Micro Devices, Texas Instruments, and Motorola provided evaluation boards.

Special thanks go to John McKean (AMD, Canada), Dan Mann (AMD, USA), Pat Eichenseer and Paul Teich (AMD 29K-hotline, USA), Hugh McGugan and Wei Chen (Motorola, Canada), and Phil Evans (Texas Instruments, Canada). L.E. Turner (University of Calgary) suggested the use of the LDI filter and allowed use of Digicap to generate the scheduling for the IIR filters. Thanks to L.T. Bruton (University of Calgary) for his suggestions regarding the manuscript. Final thanks to S. Zaman and C. Lau (NSERC summer students), who experimentally checked many of the RISC benchmark timings; E. Lok, who typed the numerous early manuscript drafts; and the reviewers.

References

1. B. Schweber, "DSP Architectures for Signal Processing Applications," *Computer Applications J.*, Feb./Mar. 1991, pp. 12-20.
2. G. Morrison et al., "Indoor Digital Radio Propagation Studies Using IDFT and ARMA Modeling," *Proc. Pacific Rim Conf.*, Apr. 1991.
3. M.R. Smith et al., "Application of Autoregressive, Moving Average Parametric Modeling in Magnetic Resonance Image Reconstruction," *IEEE Trans. Medical Imaging*, Vol. MI-5, No. 3, Sept. 1986, pp. 132-139.
4. M.R. Smith et al., "A Quantitative Comparison of the TERA Modeling and DFT Magnetic Resonance Image Reconstruction Techniques," *Magnetic Resonance in Medicine*, Vol. 19, No. 1, May 1991, pp. 1-19.
5. D.K. Mitchell et al., "The Use of Band Selectable Digital Filtering in Magnetic Resonance Image Enhancement," *Mag. Res. Med.*, Vol. 9, No. 3, Mar. 1989, pp. 353-368.
6. M.R. Smith et al., "A Hardware Implementation of an Autoregressive Algorithm," *Measurement Sci. and Technology*, Vol. 1.
7. R.O. Parker, *Am29027 Handbook, Arithmetic Accelerator for the 29K Family*, Advanced Micro Devices, Austin, Texas, 1989.
8. B.D. Green and L.E. Turner, "Digital LDI Ladder Filters Using Lattice Equivalents and Wave Concepts," *IEEE Trans. Circuit Systems*, Vol. 37, No. 1, Jan. 1990, pp. 133-135.
9. C.S. Burrus and T.W. Parks, *DFT/FFT and Convolution Algorithms—Theory and Implementation*, J. Wiley, Toronto, 1985.
10. G. Sohie, "Implementation of Fast Fourier Transforms on Motorola's DSP56000/DSP56001 and DSP96002 Digital Signal Processors," application note, Motorola, Phoenix, Ariz., 1989.
11. J.A. Booth, "RISC Compiler Optimizations," *Proc. FusionAMD Conf.*, Advanced Micro Devices, Austin, 1992, pp. 3-12.
12. J.A. Booth, "Superscale Compiler Optimizations," *Proc. FusionAMD Conf.*, Advanced Micro Devices, 1992, pp. 15-27.
13. *User's Guide, Second Generation TMS320*, Texas Instruments, Houston, Texas, 1989.
14. A. Lovrich and R. Simar, Jr., "Implementation of FIR/IIR Filters with the TMS320C10/TMS320C20," Digital Signal Processing Applications with TMS-320 Family, Vol. 1, Texas Instruments, Houston, 1986, pp. 27-47.
15. P. Papamichalis, "An Implementation of FFT, DCT and Other Transforms on the TMS320C30," *Algorithms and Implementations*, Vol. 3, Texas Instruments, Houston, 1990, pp. 53-136.
16. *DSP96002 User's Manual*, Motorola, Phoenix, 1989.
17. *i860 Microprocessor Family Programmer's Reference Manual*, Intel, Mt. Prospect, Illinois, 1991.
18. M. Atkins, "FAST Fourier Transforms in the i860 Microprocessor," Application Note Ap-435, Intel, Mt. Prospect, 1991.
19. P. Papamichalis and J. So, "Implementation of Fast Fourier Transform on the TMS320C20," Digital Signal Processing Applications with TMS320 Family, Vol. 1, Texas Instruments, Houston, 1986, pp. 69-92.
20. M.R. Smith, "Is the RISC Processor a FIR Bearing Animal?" conditionally accepted for publication in *Computer Applications J.*
21. *SPARC RISC User's Guide*, Cypress Semiconductor, San Jose, Calif., 1990.
22. M.R. Smith, "To DSP or Not to DSP?," *Computer Applications J.*, No. 28, Aug./Sept. 1992, pp. 14-23.
23. M.R. Smith, "FFT—fRISCy Fourier Transforms?" to be published in *Microprocessors and Microcomputers, 1993.*
24. *TMS320C3X User's Guide*, Texas Instruments, Houston, 1990.
25. *Am29050 Microprocessor User's Manual*, Advanced Micro Devices, Austin, 1991.
26. D. Wilson "Is RISC or DSP Best for Your Application?" *Computer Design,* Vol. 31, No. 4, Apr. 1992, pp. 65-72.

Michael R. Smith is a professor in the Department of Electrical and Computer Engineering at the University of Calgary, Canada. His research interests are in hardware and software DSP applications in imaging. He is involved in the use of autoregressive moving average (ARMA) modeling to improve the resolution, efficiency, and quality of MRI imagers used in the medical and petroleum areas.

Smith received his BSc in physics from the University of Hull, UK, his PhD in physics from the University of Alberta, Canada, and his sense of humor from his father. He is a member of the IEEE, SMRM, and SMRI, and has published more than 40 papers and application notes.

Direct questions about this article to the author at the Department of Electrical and Computer Engineering, University of Calgary, Calgary, Alberta, Canada T2N 1N4; or via e-mail at smith@enel.ucalgary.ca.

Development of an ASIC Set for Signal Processing

Our approach to signal acquisition, digitization, and processing of low-frequency physiological signals uses a chip attached to a transducer through a digital wire placed at the sensing point. The wire transmits digital information instead of an analog signal to an ASIC signal processor. This Digital Wire/Visp chip set, designed at UTSA, produces a noise-immune signal processing system usable in a variety of biosignal processing needs. We demonstrate the concept using the Visually Evoked Potential (VEP) measurement system.

Parimal A. Patel

Hemal N. Kothari

James F. Robb

University of Texas at San Antonio

Estimating sensory thresholds requires accurate and reliable tools that operate at a high confidence level. If we can reduce the effects of noise and nonlinearity problems, we can produce tools that reliably interpret wide variations of sensory signals. With this in mind, we developed a new approach to signal acquisition and processing and demonstrated it with the Visually Evoked Potential (VEP) measurement system.

Our approach involves the development of IC chip sets that can be attached to a transducer. Each set consists of a hybrid chip called Digital Wire[1] that sends digital information instead of an analog signal to an ASIC digital signal processor called Visp.[2] Digital Wire brings the analog sensing capabilities of the VEP signal, conditioning, and digitizing circuitry to the point of sensing. The advantage is a noise-immune system whose discrete output can be easily processed by digital filters to eliminate any existing nonlinear amplitudes and phase shifts.

The VEP system provides a method for estimating sensory thresholds. For example, its use in research of the human eye's response to tissue in a damaged eye can lead to solutions that protect pilots' eyesight.[3] VEP systems are particularly well suited for estimation of visual acuity before and after laser damage in noncommunicative subjects. This type of damage is a relevant and timely topic for research, given the expansion and diversified use of lasers in a variety of applications.

We designed the Visp chip to provide near-real-time processing and the complete control of up to 16 separate Digital Wire chip inputs, including one reference Digital Wire input. Such a system has applications in the US Air Force and in pediatric care.

- *US Air Force.* A pilot operating a fighter plane sometimes loses control over the aircraft during sharp turns or dives, which could cause an accident resulting in loss of the pilot and/or aircraft. To reduce such fatal accidents, the Air Force trains pilots for several man-hours—an expensive, necessary process. It would be helpful if the aircraft could be placed on automatic control (via an on-board computer) should a pilot lose consciousness. Our chip set provides this information to the on-board computer upon detecting a loss of consciousness.
- *Pediatric care.* During pediatric care, doctors often find it necessary to test a child's vision and detect any loss of vision coordination. The currently employed mechanism often fails to detect loss of coordination.

0272-1732/92/1200-0024$03.00 © 1992 IEEE

Through improvements in our chip set and additional training, we could solve this problem.

These problems can be tackled with the proposed system of digital wires and Visp. The system, connected in a mesh form as shown in Figure 1, is placed on the scalp. The system senses and processes the VEP signal, whether transient or steady-state. Of course, placing the transducers in a mesh form requires a coordination of signal data acquisition among the electrodes and additional control and communication circuits on chip. Visp provides the necessary control and communication circuits.

This data acquisition and processing system differs from other existing systems in two ways: Our system is much smaller than the other systems and emulates differential-mode signal processing by using one transducer as a reference signal. The reference signal (on an ear lobe in Figure 1) is sensed by a transducer, whose signal is sensed, processed, and digitized by a digital wire. The referenced digital wire is connected as channel 0 to Visp. Visp emulates differential-mode configuration by subtracting the reference channel's digitized signal from other active channels' digitized signals.

Digital Wire

The Digital Wire development is shown in Figure 2. The VEP signal, received from a gold-cupped transducer attached to a person's scalp, transforms a ±1-mV bipolar signal into a unipolar range. The voltage is shifted by +1 mV, and a two-stage amplifier further boosts the signal by an factor of 1,500 to the 0-3V range. The amplified voltage is further filtered through a five-stage low-pass filter with a 100-Hz cutoff frequency and a notch filter to eliminate 60-Hz noise. The filters are implemented using switched capacitors.

A sample-and-hold circuit permits the filtered output to be sampled and held for more than 1.0 ms (> 16 bits×52 μs/bit) to ensure nearly stable voltage as one of the inputs to a comparator of the 14-bit analog-to-digital subsystem. The other input of the comparator is received from the output of the digital-to-analog converter (DAC) subsystem, consisting of a successive approximation register (SAR). The Digital Wire samples the signal in common mode, and Visp calculates it in differential mode through subtracting a digitized reference signal.

We found it necessary to increase the precision of the sampling and calculating process. We selected a 14-bit design to achieve a resolution of 2 mV/16,384 = 0.12 μV. The SAR output serves as switch control input in the DAC. The DAC can be implemented using either a resistive or switched capacitor network. However, due to poor resistance tolerance in the CMOS layout and the fact that a better than 0.1 percent accuracy can be achieved in capacitor ratio, we chose to realize the DAC using a switched-capacitor charge redistribution configuration. Conversion completes in 16 clock cycles

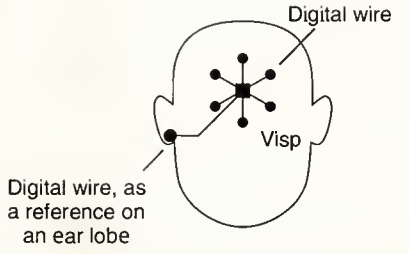


Figure 1. A complete VEP system.

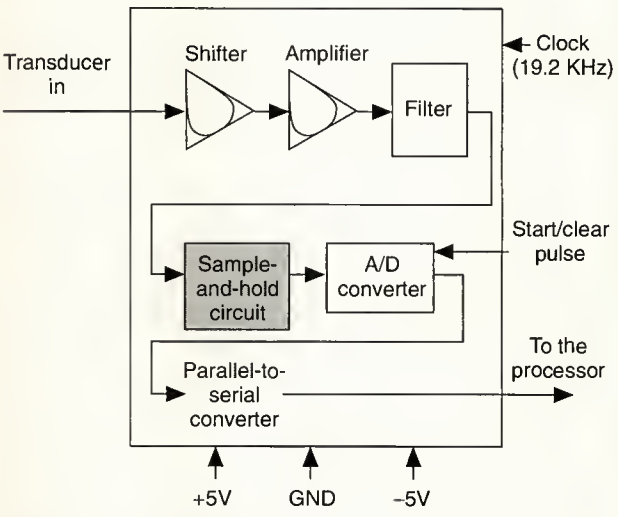


Figure 2. The Digital Wire.

(1 start bit, 14 data bits, and 1 stop bit) after the start pulse enters the chip. In designing the complete Digital Wire, we started from the basic gates, at the integrated device level.

Operational amplifiers and shifters. One of the most important circuits in analog circuit design is the operational amplifier. It primarily provides sufficient gain to define and implement analog signal processing functions through the use of negative feedback. Such analog signal processing functions include amplification, integration, and summation.

The steps in designing an operation amplifier depend on the desired values of parameters.[4,5] Target specifications of a CMOS operation amplifier designed to drive an on-chip capacitive load are a 60-db differential gain; 1-MHz unity gain bandwidth; ±1V/μs slew rate; and input differential impedance of 10^{12} ohms. The need for a second stage for the CMOS operation amplifier is more obvious since the gain of the first stage is not sufficient and the output resistance is very large, which is not suitable for low resistive loads.[6] The second stage reduces the output resistance, increases the overall volt-

age gain, and increases the output swing. To reduce the input offset voltage, we laid out the operation amplifier in a centroid configuration.[7] Figure 3 shows the two-stage operation amplifier designed to drive the on-chip load. Figure 4 shows the open loop gain plot of the operation amplifier (SPICE simulation result). A tracking gain bandwidth characteristic curve appears in Figure 5.

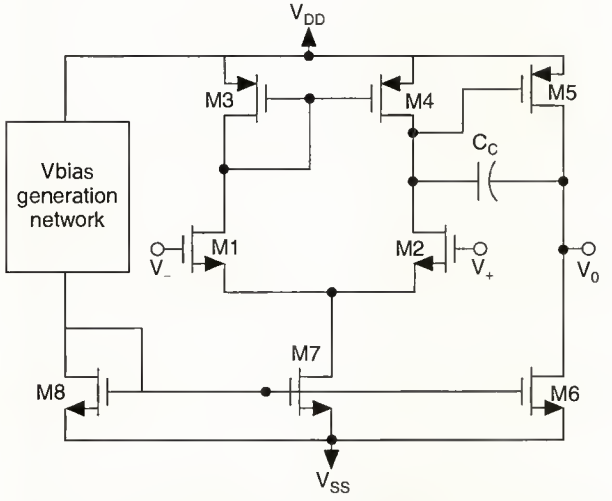


Figure 3. Two-stage operation amplifier.

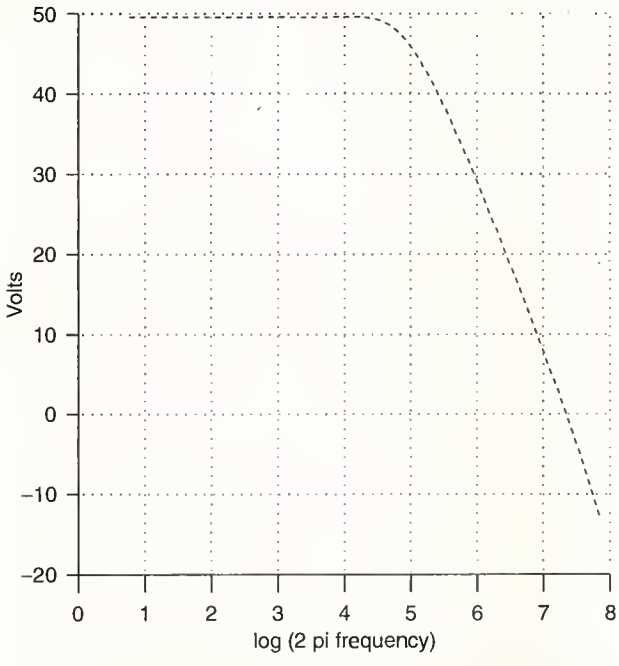


Figure 4. Open-loop gain plot of the operation amplifier (SPICE simulation result).

Because of the bipolar nature of the signal, a shifter is designed so that the signal range can be transformed to a unipolar range. So we used the shifter configuration that permits the designed operation amplifier to differ only in the length and width of one pair of NFET (N-channel field effect transistor) and PFET (P-channel FET) transistors to add offset at the input stage. This design permits the original signal to be shifted from ±1 mV to 0-2 mV.

The signal can be easily and accurately processed since the shifted signal can be amplified from 0-2 mV to a suitable voltage level of 0-3V. Two cascaded CMOS operation amplifiers achieve the needed amplification factor of 1,500. An inverting amplifier configuration using a feedback resistor would be a simple choice for this purpose. However, the amplification factor is a function of $R_{feedback}/R_{in}$. Because of poor resistor layout tolerances (±25%), we decided on a configuration using switched capacitors, which produces a high (0.1 percent) capacitor ratio. Figure 6 shows the switched-capacitor amplifier circuitry.

Filters and sample-and-hold circuit. Due to the information-bearing nature of the input signal, the preservation of its original shape is of prime importance. Mainly, we needed a low-pass filter that would block all high frequencies (typically, all frequencies above 100 Hz) so that we could study the signal of interest. The second requirement was to eliminate the power line frequency (60-Hz noise), and for this purpose we needed a notch filter.

We decided therefore to use a ninth-order Chebyshev filter, which is a cascade of low-pass and notch filters. Thus, we can obtain optimal performance and a major savings in silicon area. The particular design we've implemented has the minimum number of stages when compared to any other configuration possible and also the least value in terms of components used, thus saving silicon area.

We selected a 0.1-dB passband deviation and a switched capacitor biquad topology. We used Filter Designer, a tool for Pspice Circuit Synthesis,[8] and then later tried using the Switcap software simulator.[9] After checking results of the Spice3e[10] raw file (which is generated as a netlist by Switcap), we confirmed the final design. The filter also includes an equalization network that takes care of the phase delay.

The sample-and-hold circuit samples the input when the sample (start) pulse goes low and holds the output when the start pulse goes high, using the operation amplifier, capacitors, and CMOS switches. It is important that this circuit can rapidly track changes in the input voltage when in the sample mode and not discharge the capacitor when in the hold mode. The sample-and-hold circuit samples at a1,000-Hz rate.

The dynamic performance of the converter depends largely on the dynamic characteristics of the operation amplifiers

and comparators. Therefore, the slew rate, settling time, and overload recovery time of these circuits are important.

Comparator. The comparator is noninverting as the output moves from the low state to the high state when the voltage V_p becomes larger than V_N. The performance of a comparator can be characterized by

- its resolving capability or threshold sensing,
- the input offset voltage,
- the speed or propagation delay time, and
- the input common-mode range.

A comparator can be implemented using three methods: a high-gain differential amplifier, a positive feedback, and charge balancing. The key attribute of the differential amplifier is its ability to amplify the difference between the inverting and noninverting inputs over a wide common-mode range. As a result, the threshold point or trip point can be independent of the process and supply voltage variations to a first-order approximation.

The input offset voltage of the differential amplifier results from the mismatches in the devices. Mismatches of this type are unavoidable and caused by imperfections in the process. Offsets can be minimized by using the common centroid geometrical layout. It is desirable to keep the number of bends in the layout for the two devices the same. The input offset voltage can also be reduced by using large areas for the devices and by keeping the gate-source voltage small.

The output pole sets the propagation time of the differential amplifier used as a comparator, and the propagation time is determined by the large signal response. The load capacitance seen by the differential amplifier comparator greatly influences the propagation delay time. The gain of most CMOS differential amplifier comparators is too small to give satisfactory resolving capability. To increase the gain, we used the just-described, two-stage operation amplifier for the CMOS comparator.

The comparator compares the output of the DAC (which is the current approximation to the analog signal) to the output of the sample-and-hold circuitry (which is the conditioned signal being converted). The resulting signal DCP resets the current approximation bit, if the signal being converted was less than the approximation signal.

Digital-to-analog converter. As seen in Figure 7, next page, a successive approximation register[11] forms an integral part of the DAC. The SAR's basic function is to generate the 14-bit digital word representing the current input voltage from the sample-and-hold circuitry. This is accomplished through 14 approximations in which the 14 bits are set and reset in succession. The bits are reset as necessary in response to a comparison with the current input sample. The final digitized result is output (every clock cycle) by the storage register, until all 14 bits shift out.

The SAR consists of three main parts:

- the state machine and decoder,
- the storage register, and
- reset logic.

The state machine is a 4-bit counter whose output appears in gray code; that is, only 1 bit changes between any two given states. This design eliminates the static hazards present in a normal 4-bit counter. The state machine output is decoded using 16 NAND gates to produce high-level logic in response to each successive state. The outputs of the de-

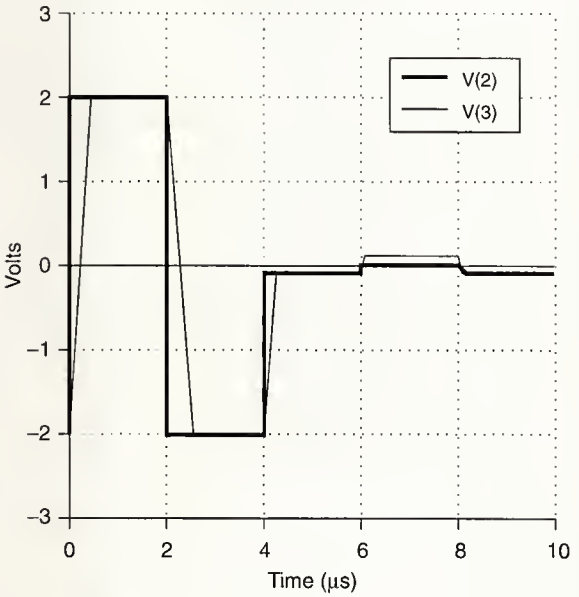


Figure 5. Tracking gain bandwidth characteristic curve.

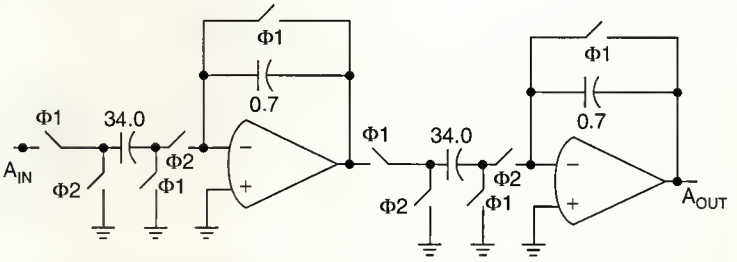


Figure 6. Two-stage switched-capacitor amplifier circuit.

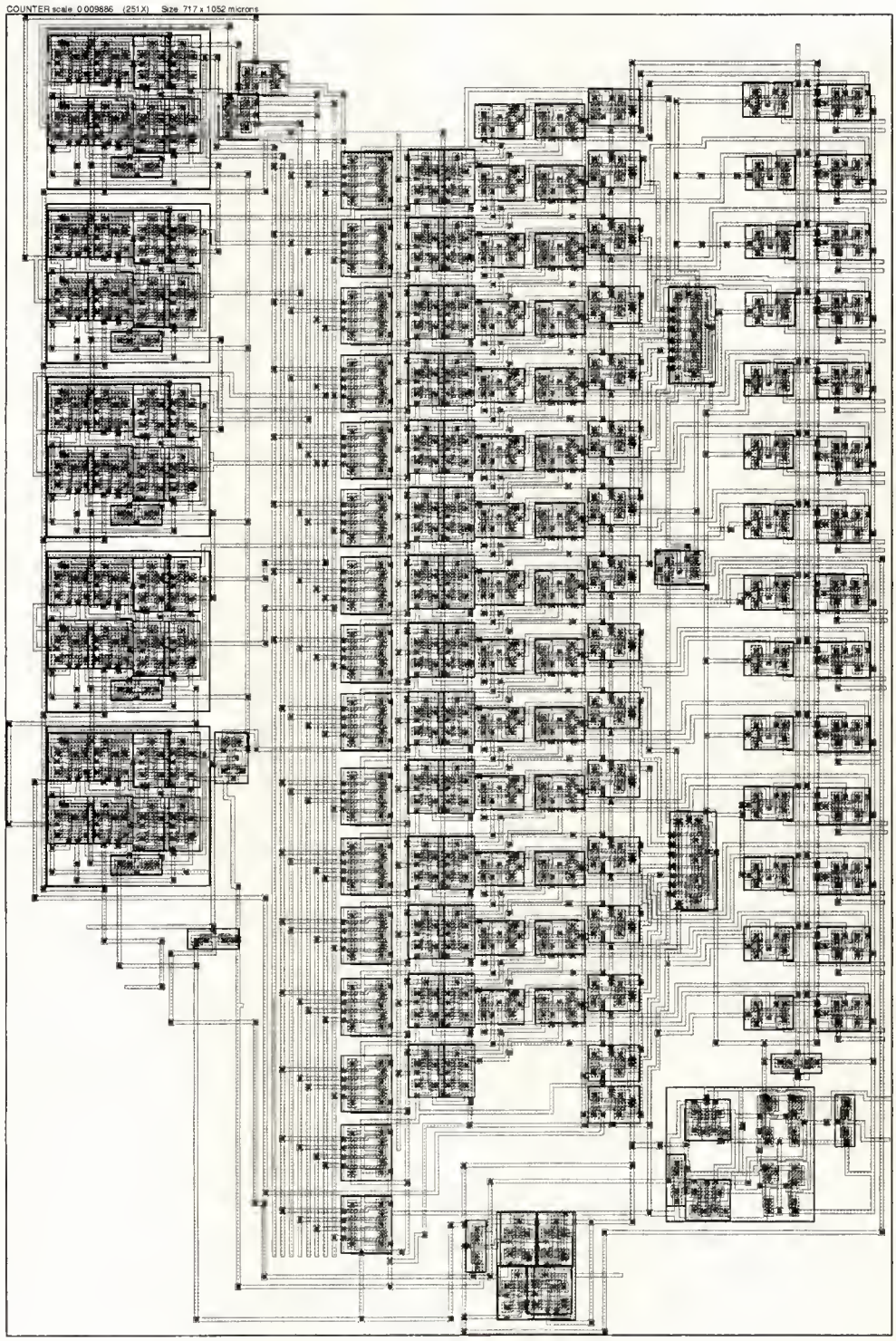


Figure 7. A successive approximation register (SAR) layout.

coder are connected to the set inputs of the 14 SAR latches, which comprise the storage register, thus setting each bit in succession. To permit the least significant bit to be reset, we allowed the state machine to return to the initial state before the clock is stopped, thereby clearing the set pulse to the least significant bit latch. The output of the storage register feeds into the DAC. The storage register output generates the two-phase clock pulses, which are used in the DAC section to charge or discharge the capacitors of the switched-capacitor analog-to-digital converter network.

The output of the storage register feeds back to the DAC, whose output is compared to the current signal from the sample-and-hold circuitry.[5] If the current approximation is greater than the present sample, a reset pulse is generated at the next clock pulse to reset the corresponding bit. This cycle repeats with each subsequent approximation (from the most significant bit to the least significant bit), resulting in increased accuracy with each step. The final result should be accurate to ±1 least significant bit. The entire cycle takes 16 clock cycles.

Digital-to-analog conversion is simpler to implement and can be done by three methods: charge-balance, resistive voltage division and a combination of the two, and serial DAC. We chose charge-balance DAC for this design.

The most simple DAC works in one stage on the charge redistribution principle and is binary-weighted. It uses capacitors, switches, and an operation amplifier as a buffer. The first step in a conversion discharges all capacitors during the Φ1 phase. During the Φ2 phase period the binary switches are closed or opened, depending on whether the bit is a 1 or a 0.

At this point an equivalent circuit for this converter is simply a capacitive attenuator. As an attenuator, some or all of the capacitors may be connected to V_{ref}. Typically, the bottom plates of the capacitors connect to the binary switches, and the top plates are connected in common. Note that the bottom-plate parasitic effect is negligible in this configuration, and through the charge-balancing technique the influence of the top-plate parasitic can be minimized. The accuracy of the capacitor and the area required are both factors that limit the number of bits used. The ratio for MOS technology capacitors may be as good as 0.1 percent. A resolution of 10 bits requires a capacitor ratio of 1,024, which would require too much area. These considerations led to the development of a three-stage DAC to optimize silicon area usage. Figure 8 shows the schematic of the 14-bit DAC with unipolar V_{ref}.

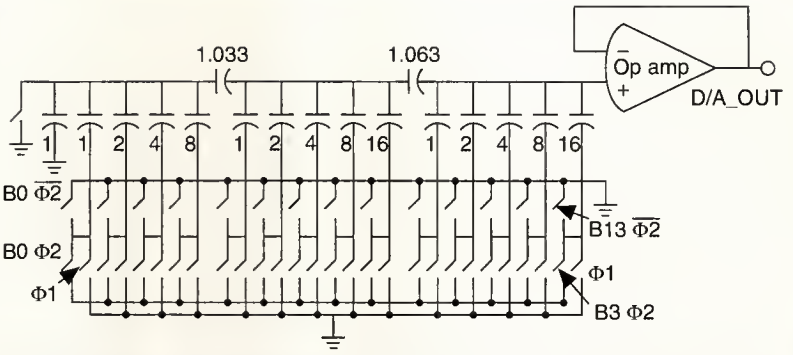


Figure 8. A 14-bit digital-to-analog converter. All capacitors shown in pF.

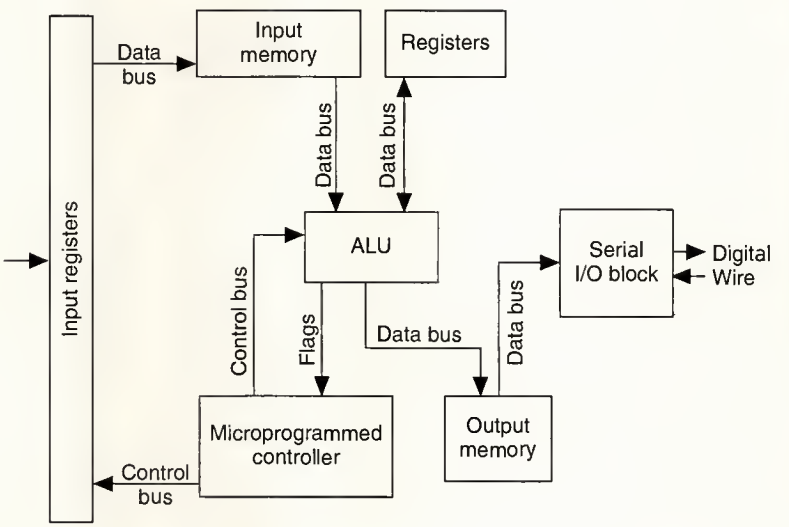


Figure 9. Visp block diagram.

Visual Signal Processor

Visp allows the complete control and filtering of digital signals provided by the Digital Wire. The Visp's application-specific IC (ASIC) design allows for a new method of acquiring, processing, and storing VEP signals. The output of this device could be used to drive a variety of control equipment in near real time. It maintains a programmable environment that the user may modify easily from a remote location. Figure 9 depicts the Visp block diagram.

The purpose of the Visp device is to control the data acqui-

sition section of the VEP system by generating a data conversion pulse. Visp must also control the flow of measurement data through the system, and allow on-chip digital signal processing (digital filtering and/or transformation) of the measured VEP data. The input section of this device takes up to 16 separate digital input streams from the VEP acquisition chips, processes the data according to preprogrammed filters, and then outputs the processed data via an RS-232 serial line.

Visp contains many separate sections that perform different tasks concurrently. It contains 16 input channels, 8 Kwords of 16-bit each data memory, a fully functional 16-bit ALU, programmable registers, a microprogrammed controller, a serial I/O channel, and a clock generator. The inclusion of all these devices on the chip lowers the external devices needed for support to a minimum.

Basically, the device provides clocking pulses and conversion pulses to the Digital Wire devices, and receives simultaneous inputs of up to 16 separate digital measurements. These measurements are read from the input registers on the Visp into the on-chip memory for storage and further digital signal processing. Upon completion of the signal processing, processed data are sent through a serial I/O channel to a remote computer.

The clock generator and the microprogrammed controller work in tandem to provide the required clocking signals to the VEP acquisition system. The clock generators take an input system clock of 1 MHz, divide it into the required 19.2-kHz clock necessary for serial communication, and further divide it into a 100-Hz clock for the VEP acquisition start pulse. Measurement data are input into one of the sixteen 16-bit input registers on the Visp. Data then transfer to on-chip memory over a 16-bit data bus for storage. The microprogrammed controller and 16-bit ALU apply the digital signal processing to the stored data based on programmed filter characteristics. Both digital and analog components transfer data serially, and the 19.2-kHz clock pulse and special registers permit serial communication to take place.

The Visp clock circuitry requires a 1-MHz TTL clock to generate internal and external timing. Necessary timing pulses include the 19.2-kHz clock (CLK_SERIAL) used by both the Digital Wire and Visp chip for serial communication. Also, Digital Wire needs a 100-Hz conversion pulse (CLK_CONVERT) to start VEP measurement.

The signals are generated by basic division of the system clock by predetermined values, using full 8-bit binary count-up counters with full preset, clear, and load capabilities. Counter reset occurs upon completion of one count by basic logic comparisons to a stored count value held in the clock register using NAND and NOR gates. A 16-bit parallel-input/parallel-out (PIPO) register holds count values (52 and 192) for both 8-bit counters. When the output of the counters matches the output of the register, the system resets the counters and sends the clock pulse to the other on-chip systems and output pins for use by the Digital Wire.

The voltage divider is one of only two totally nondigital sections necessary on the Visp chip. To provide necessary voltages for on-chip components and especially serial communication requires production of a variety of voltages. Three off-chip inputs of +12V, −12V, and ground create the +5V, +3V, and −3V signals on chip. The +5V supply device (V_{CC}) and ground (GND) power, while the remaining voltages are used entirely for serial communication. The +3V and +12V receive and transmit logical 0 communication bits, while the −3V and −12V are used for logical 1 bits. A voltage divider generates the required voltages, tapping them off the +12V to ground inputs and off the −12V to ground inputs.

Serial communication. The Visp serial communication section breaks down into six separate subsections: two for receiving data, two for transmitting data, and two for communication logic.

A 10-bit serial-input/parallel-output (SIPO) register and an RS-232-to-TTL voltage converter receive data. The 16-bit register is clocked by the 19.2-kHz clock whenever data arrives at Visp from an external source. We designed the receive function so it could be programmed by users or mode setting on the Visp. When 10 bits are in the SIPO register, including the one start and one stop bit, the data byte is latched onto the data lines and stored in nondata sections of output memory, while the start and stop bits are checked for validity by the ALU. Communication with the Digital Wire takes place at TTL levels and communication with the distant controller at the RS-232 level.

A 10-bit parallel-input/serial-output (PISO) register transmits data. The 16 data bits representing a valid VEP measurement word are taken from output memory and broken into two 8-bit bytes. Then these are combined with the necessary start and stop bits to communicate at 19,200 baud, no parity, 8 data bits, 1 start bit, and 1 stop bit (19200,N,8,1,1). Each byte is sent individually (high byte then low byte) by placing it in the 10-bit PISO register with start and stop bits.

The communication logic section has one input and output line for clear-to-send and data ready events. The input line is a clear-to-send signal for the Visp to activate when it is ready for the host to send data; it would remain in this active state unless busy. The output line indicates to the host system that data are ready to send to the host.

Registers and data storage

The VEP input section of the chip contains 16 registers. These registers are all 16-bit SIPO registers similar to the serial transceiver receiver section. Because the digital wires are near the Visp, communication between the devices takes place at the TTL level. Their purpose is to input the acquired measurement word from the VEP acquisition devices. Sixteen SIPO registers can capture 16 separate measurements simultaneously.

Each register permits 16 bits of information to be clocked

in sequentially at the conversion rate. As the digital wires produce one bit of output per clock based upon their successive approximation register, the Visp will input that data simultaneously. On completion of transmission, the 16 data bits are placed onto the data lines and stored in input memory.

Storage of measurement data by Visp is essential in the process of applying digital signal processing to the data before transmitting a final result. We provided an 8K×16 static RAM to store input data and output results. Input memory stores new data, while output memory stores processed data.

The Visp device uses a 12-bit address bus to address both sections of the memory. A 12-bit PIPO register called MAR stores current memory addresses, while a separate MAR stores both input and output memories.

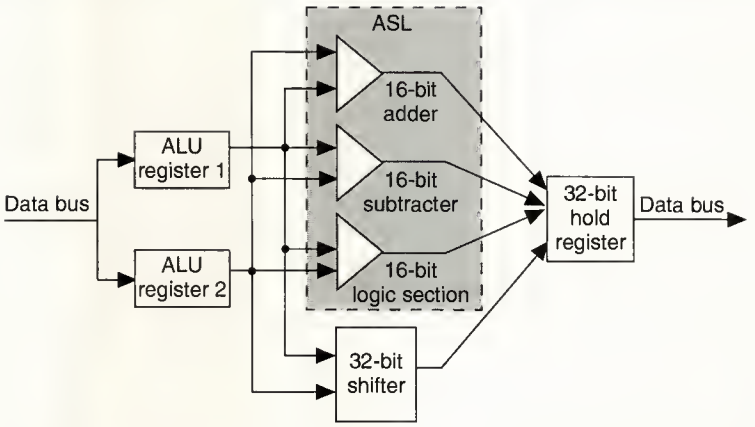


Figure 10. ALU block diagram.

ALU and microcode controller. The 16-bit ALU is a fully functional unit capable of addition, subtraction, shifting, and logic functions. Inputs to the ALU are stored in two 16-bit registers and fed into a full 16-bit adder/subtracter/logic (ASL) section (see Figure 10) and a 16-bit shifter section. Depending on the control signals sent by the microprogrammed controller, specific functions are performed on the data resulting in flags being set and the result being stored in the ALU output registers.

We designed the ASL section to use four 4-bit, binary, full adder/subtracters along with carry-lookahead circuitry; we also combined the adder/subtracters in serial to provide carry-ripple-through capability. Flags set by the ASL section include equal and carry, two of the five flags used by Visp. A zero flag is set by NORing all the ASL outputs.

The ALU also uses a 16-bit shifter capable of bidirectional 1-bit shifts. Depending on a control word, the shift is only 1 bit in either direction, with the saved bit being stored in the shift_left or shift_right flags. The multiplier uses these flags for determining partial product addition.

Multiplication and division are both programmed into the controller for use in the digital signal processing algorithms. Up to two 16-bit words can be multiplied, producing a 32-bit product that would be stored in the ALU output registers. Division takes place similarly, using microcode instructions; after division, a 16-bit quotient and remainder are stored in the two 16-bit output registers. Although not used by the median filter algorithm, design of multiplication and division into the microcode instruction set will allow future programmability.

The final and most complex section of Visp is the microcode controller. The controller's many different parts include the microcode memory, microcode counter, microcode address decoder, jump logic, instruction register, instruction decoder, microcode buffer, and vertical microcode decoder. We designed the microcode store as a 60-bit-wide ROM with an addressable size of 1,024 words. The microcode is in partial horizontal format to optimize the speed and storage space.[12]

Inputs to the controller include reset and enable signals from external pins, the 1-MHz system clock, flag registers, and data lines. Upon a reset, the microcode counter is reset along with the program counter. The microcode counter operates only when the enable line is active, meaning Visp will be idle when the enable line is disabled. Normal operation includes a fetch cycle followed by a decode and execution cycle. Instructions are decoded after they are fetched and placed into the 16-bit instruction register. The microcode counter is loaded with the instruction value corresponding to the location of microinstructions in the microcode memory. The microinstructions execute sequentially by being loaded into the microcode buffer followed by either decoding or by conditional jump logic. Conditional jump logic is applied when a micro jump condition exists. If no conditional microcode jump condition exists, the decoded microcode is sent to the remainder of the device in the form of control lines.

Signal processing

One of the main purposes for the development of Visp was to provide on-chip and near-real-time digital signal processing capability along with measurement data control. The on-chip need for a variety of user-programmable filters led us to design a median filter algorithm into the device with a variable size mask, depending on the programmed window size, found in the status register.

Although the median filter is designed into the microcode, flexibility lies with the user/programmer who can create user-specific filters and store them in the on-chip control memory. The microprogrammed controller controls data and, depending on the number of active channels programmed into the

status register, efficiently uses the memory to create multivariate data storage for filter algorithms.

Data retrieved from the digital wires may take many different forms, depending on the user-selectable factors—the most important of which are the number of active acquisition devices and the conversion pulse rate. Connection to only two acquisition devices yields a single-dimensional voltage intensity function, in which each voltage value will be a given time quanta (1/conversion pulse rate) separated at 14-bit resolution. Active site data are subtracted from the reference data to obtain the differential information of interest. Having time-related input data (for example, eye stimulus) allows the user to correlate the output potential to the input, a data type referred to as one-dimensional potential data.

The other data type capable of being stored and processed by Visp is multidimensional input data coming from up to 16 separate acquisition devices (one reference, up to 15 other active sites). These data are referred to as multidimensional potential data. Each active site's data are subtracted from channel 0, the reference signal. Multidimensional potential data are more suited for image processing functions (median filters, order statistic filters, and so on) being applied to it than potential data.

Storage is a key problem when dealing with sampled data. The 16-bit memory depth allows for a 15.26-μV resolution on a 1V scale; however, this use takes up memory very quickly. Also, with multi-input capability designed into Visp, memory for data storage becomes even more limited when up to 16 samples are taken on every conversion pulse. With approximately 4,096×16-bits of data storage in input memory, a maximum of 4,096 voltage potential samples could be stored and processed. Given a rate of 100 samples per second, this permits a maximum recording of 40.96 seconds of data from one active channel or up to 1.50 seconds for up to 15 active channels.

Processing algorithms

The user can apply many different forms of median filters to the sampled data. If 1D data is collected, a 1D median filter may be applied to the data. This filter uses a 1D mask with a window width from three, five, seven, or nine data points. The size of windows can be programmed in the status register. We chose median filtering as the primary filter due to its image-enhancing capabilities. Using a n-window mask on sampled 1D data has very beneficial effects in smoothing spurious effects or spikelike components caused by external noise. Additional benefits include the retaining of edges—or in the 1D case, sharp contrast. Median filtering in the multidimensional case is even more beneficial. Users can mask sizes of 3×3, 5×5, 7×7, or 9×9 in the status register.

Visp applies median filtering based upon a precoded algorithm stored in microcode. This routine takes the programmed mask size found in the status register, and using the complete data set, sorts each mask worth of data into a list using the ALU's comparison capability. Following the sorting of three to nine words in the 1D case or nine to 81 words in the multidimensional case, the median value is determined and stored in output memory reserved for the filtered image. When complete, the image is transmitted to the host system for display or further analysis.

THE VEP MEASUREMENT SYSTEM, consisting of the digital wires and Visp, brings about a new approach to the field of biomedical data measurement, control, and processing. The hybrid Digital Wire reduces or eliminates problems associated with several disjoint components, connected through long cables. Visp's programmability and flexibility enable the constant increase in capability for the device. With a complete on-chip microprogrammed environment, the system can maintain total control of up to 16 separate channels.

We considered only the median filter as a preprogrammed filter; however with the programmability aspect of Visp, other image enhancement and processing techniques can be implemented. Some of these algorithms include both frequency and spatial domain methods. Median filtering is a form of spatial domain processing that could be modified very easily into neighborhood averaging or some other windowing scheme. Some future filter techniques to consider are neighborhood averaging, histogram equalization, FFT, and the Hough transform.

The small size and relative cost per device show that VEP can be used in lieu of more expensive and advanced general-purpose computers. Although the current development applies to a specific biosignal acquisition and processing application, the VEP system can be extended to other physiological signal acquisition, processing, and interpretation applications.

Acknowledgment

A US Air Force School of Aerospace Medicine In-House Laboratory Independent (ILIR) grant partly funded this work.

References

1. H. Kothari, "Development of Integrated Electrodes to Measure Visually Evoked Potential, master's thesis, Univ. of Texas, San Antonio, College of Sci. & Eng., 1992.
2. J. Robb, "Visually Evoked Potential Digital Signal Processor (Visp)," master's thesis, Univ. of Texas, San Antonio, College of Sci. & Eng., 1992.
3. D. Regan, *Human Brain Electrophysiology*, Elsevier Science Publishing Co., New York, 1989.

IEEE MICRO

Reader Survey

HOW ARE WE DOING? We want to serve you better; tell us how by completing and returning the following by mail or fax to (714) 821-4010.

1. I read *IEEE Micro*:
- ❑ always
- ❑ often
- ❑ sometimes
- ❑ never

2. I find its scope of coverage:
- ❑ too narrow; should add areas of _____
- ❑ just right
- ❑ too broad; should cut areas of _____

3. Specific topics I would like to see covered in *IEEE Micro* next year are:

1. _____
2. _____
3. _____

4. I find the articles in *IEEE Micro*:
- ❑ are practically oriented
- ❑ are too theoretical
- ❑ have technical value
- ❑ point to new findings
- ❑ are too long
- ❑ need longer abstracts (now 50-75 words)
- ❑ need signposts on each page to highlight the content

5. I rank departments as follows: (circle appropriate number)

	Never read=1				Always read=5
Micro Law	1	2	3	4	5
Micro Review	1	2	3	4	5
Micro Standards	1	2	3	4	5
Micro News	1	2	3	4	5
New Products	1	2	3	4	5
On the Edge	1	2	3	4	5
Software Report	1	2	3	4	5

6. If I could change anything about *IEEE Micro*, I'd:

_____

_____

_____

7. My issue is delivered:
- ❑ on time
- ❑ in good condition
- ❑ late
- ❑ damaged

8. I also read *IEEE Design & Test of Computers*:
- ❑ yes
- ❑ no

Other journals I read are: _____

_____

_____

9. My job title and application area are: _____

_____

_____

10. I hold the following degree(s): _____

_____

11. I've submitted articles to *IEEE Micro*:
- ❑ yes
- ❑ no

12. I liked the following articles from the August 1991 to October 1992 issues: (Check as many as necessary.)

August 1991 - Far East
- ❑ Toward advanced parallel processing
- ❑ Gmicro/100 32-bit microprocessor
- ❑ ITRON-MP adaptive real-time kernel
- ❑ Sure System 2000 communications processor
- ❑ KMDS expert system: integrated hardware/software design

October 1991 - General Interest
- ❑ Simulating a function of visual peripheral processes
- ❑ Computer analysis of the myoelectric signal
- ❑ Alphorn remote procedure call environment
- ❑ RISC processor for embedded applications in an ASIC
- ❑ Performance and the i860

December 1991 - Database Machines
- ❑ VLSI accelerators for large database systems
- ❑ Associative accelerator
- ❑ Fine-grain architecture for relational database operations
- ❑ Parallel, scalable database computer
- ❑ Rinda relational database processor

February 1992 - General Interest
- ❑ Scalable coherent interface and related standards
- ❑ Unix and the Am29000
- ❑ Hardware requirements for neural network classifiers
- ❑ Experimentation with hypercube database engines
- ❑ Conformance testing of VMEbus/Multibus II products

April 1992 - Hot Chips III
- ❑ The Mips R4000 processor
- ❑ Message-driven processor
- ❑ Motorola 88110 superscalar RISC
- ❑ Proposed SSBLT standard

June 1992 - Associative Processors and Memories
- ❑ Associative processors and memories survey
- ❑ Pattern-addressable memory
- ❑ Dynamic associative processor for machine vision
- ❑ Cascading content-addressable memories
- ❑ Associative processing module for vision

August 1992 - Microelectronics in Europe
- ❑ CMOS technology trends and economics
- ❑ Consumer applications
- ❑ A/D technologies for mixed-signal processing
- ❑ Trends in library development
- ❑ European activities for EDA standardization
- ❑ Floating-point processors for parallel processors
- ❑ Special report: MITI's Real-World Computing program

October 1992 - Video Coding Hardware
- ❑ Processing hardware for real-time coding
- ❑ Digital video coding techniques for HDTV
- ❑ A DSC-HDTV video decoder system
- ❑ A 160-Mpixel/s IDCT processor for HDTV
- ❑ Programmable Vision Processor/Controller
- ❑ V.42bis standard for data-compressing modems
- ❑ Supercomputer performance for neural net simulation with an array of DSPs

Return your comments by fax to (714) 821-4010 or fold, staple, and mail this card.

NO POSTAGE NECESSARY IF MAILED IN THE UNITED STATES

BUSINESS REPLY MAIL

FIRST CLASS PERMIT NO. 38 LOS ALAMITOS, CA

POSTAGE WILL BE PAID BY ADDRESSEE

IEEE COMPUTER SOCIETY
PO BOX 3014
LOS ALAMITOS CA 90720-9804
USA

4 C.L. Britton et al., "Characteristics of CMOS Analog Integrated Structures Using MOSIS," tech. report, Oak Ridge National Laboratory, Oak Ridge, Tenn.
5. *Analog MOS Integrated Circuits*, P. Gray, D. Hodges, and R. Brodersen, eds., IEEE Press, New York, 1980.
6. R.L. Geiger, P.E. Allen, and N.R. Strader, *VLSI Design Techniques for Analog and Digital Circuits*, McGraw-Hill, 1990.
7. P.E. Allen and D.R. Holberg, *CMOS Analog Design*, Holt, Reinholt, and Winston, New York, 1987.
8. *PSPICE Documentation*, Micro Sim Corporation, Laguna Hills, Calif., 1989.
9. K. Suyama, *User's Manual for SWITCAP2 Version 1.0*, Dept. of Electrical Engineering and Center for Telecommunications Research, Columbia University, New York, 1991.
10. B. Johnson et al., *SPICE3 Version 3e1 User's Manual*, Dept. of Electrical Engineering and Computer Sciences, University of California, Berkeley, Calif., 1991.
11. R.N. Mayo et al., *Magic VLSI-Layout Tools*, DECWRL/Livermore Magic Release, Livermore, Calif., 1990.
12. M. Rafiquzzaman and R. Chandra, *Modern Computer Architecture*, West Publishing Company, New York, 1988.

Parimal A. Patel's biography and photograph appear on p. 9 in this issue.

Hemal N. Kothari is a graduate student in electrical engineering pursuing his master's degree at the University of Texas at San Antonio, where he works as a research and teaching assistant in the Division of Engineering. His research interests include application-specific integrated electrode design, ASIC DSP design, VLSI and microprocessor design, image processing, and signal processing.

Kothari received the BE (electronics) degree from the University of Bombay, India, and is a member of the IEEE and the Computer Society.

James F. Robb is also a graduate student in electrical engineering at the University of Texas at San Antonio. He works as a test and evaluation engineer for the US Air Force Electronic Warfare Center, Kelly Air Force Base, Texas. His research interests include VLSI design, biomedical instrumentation, and signal processing.

Robb received a BSEE from the United States Air Force Academy, Colorado Springs, Colorado. He is a member of the IEEE and the Computer Society.

Direct questions about this article to Parimal Patel at Division of Engineering, University of Texas at San Antonio, San Antonio, TX 78249; or via e-mail at ppatel@voyager1.eng.utsa.edu.

Reader Interest Survey

Indicate your interest in this article by circling the appropriate number on the Reader Service Card.

Low 156 Medium 157 High 158

Analog VLSI Neural Networks for Impact Signal Processing

The 80170NX Electrically Trainable Analog Neural Network recognizes objects in real time. It uses a discrete Fourier transform to preprocess an accelerometer output waveform that is subsequently recognized through a multilayer Perceptron neural network. We show that neural network hardware operating in a linear mode can perform conventional signal processing functions. The similarity of neural network computations to linear signal processing functions makes it exceedingly straightforward to integrate neural networks and conventional signal processing in the same system.

Jeff Brauch

Simon M. Tam

Mark A. Holler

Arthur L. Shmurun

Intel Corporation

Interest in using neural networks for signal processing has grown rapidly in recent years, bringing about the first IEEE Workshop on Neural Networks for Signal Processing in 1991.[1] (See the Neural Network Basics box.) The nonlinear characteristics of "artificial" neurons and the learning algorithms used to determine weights for them are the characteristics of neural networks that differ the most from conventional signal processing techniques. A typical neural network calculates dot products, the same function used in discrete Fourier transforms (DFTs)[2] and finite and infinite impulse response filters (FIRs and IIRs). In addition, the neural network transforms each dot product nonlinearly (sigmoidal output), which enables the neural network to perform certain functions that linear signal processing systems cannot.

Many DSP algorithms can proceed in a straightforward manner on a neural network combined with a tapped delay line. Figure 1 shows the similarity between an FIR and a neural network. Adding a tapped delay line to a neural network and eliminating its nonlinear transfer function produce a formal equivalent of an FIR.

When using a neural network in a signal processing application, designers want to take advantage of the large body of signal processing knowledge that already exists. For instance, if a low-pass filter function is required, it will be more efficient to use the well-known equations for an FIR filter than to try to train a neural network with a tapped delay line to behave like a low-pass filter. We followed this approach in our impact recognition system and used a DFT to preprocess an incoming waveform.

Impact recognition

Characterizing impacts, or collisions, is a natural application area for neural networks. The characterization may need to occur in real time if some action is to be taken before the object leaves contact or before some damage begins to occur. For example, automobiles need a system that can determine whether to inflate an air bag restraint within approximately 5 ms of an impact. Vibration cancellation systems may seek to cancel vibrations with the maximum frequency a human can hear, about 10 kHz. Such a system requires a response time of less than 100 µs. The fundamentally parallel neural network fits this type of application because it can compute a response very quickly when implemented in parallel hardware.

Also, a neural network can "learn" to handle complex patterns rather than being programmed.

continued on p. 36

0272-1732/92/1200-034$03.00 © 1992 IEEE

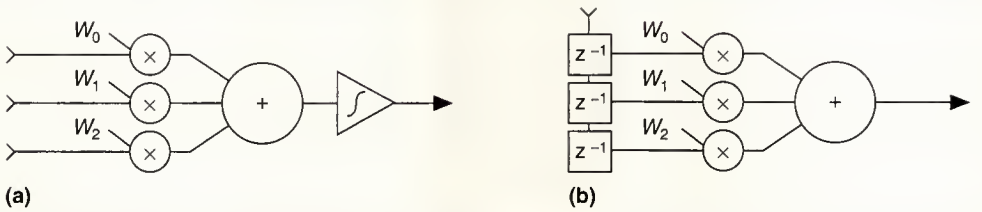


Figure 1. Similarity of an artificial neuron (a) to an FIR filter (b). *W* indicates a weight.

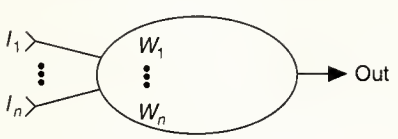
Neural network basics

Neural networks are composed of parallel, connected processing elements called *neurons*, which have a large number of inputs but only one output. Each input stores a weight inside it, as shown in Figure A. Individual inputs can be thought of as components of a vector, the *input vector*. Similarly, the weights can be thought of as a *weight vector*.

Each input is multiplied by its associated weight, and the products are added. This operation is equivalent to taking the inner product of the input vector and the stored weight vector. The result is then made *nonlinear* by mapping on a *sigmoid function* (see Figure B). If the inner product of the input and the weight vectors exceeds a specified threshold, the neuron's output is brought to a high state; otherwise it remains in the low state. The region around the threshold has a positive slope and is also known as the *sigmoid's gain*.

The neurons just described are organized in interconnected layers, as in Figure C. In a multilayered network all the layers that do not feed the outputs are called *hidden layers*, and the layer whose outputs connect to the external world is called the *output layer*. The inputs of each layer's neurons are the outputs of each of the previous layer's neurons. So, all neurons are connected to each neuron in the next layer.

Instead of being programmed to do their jobs, neural networks learn from experience. Though other methods exist, the *back-propagation algorithm* has become one of the most popular. It works by presenting a network with a set of inputs called the *training set*. Its response to each input is compared to the expected response, and the difference between the two is computed. Weights are then adjusted depending on the difference. This process is repeated over the entire training set until the network's outputs are within an acceptable error margin of the correct output.



$$\text{Out} = s\left(\sum_{k=1}^{n} I_k W_k\right)$$

where *s* is a sigmoid function

Figure A. A neuron; *I* indicates the input, and *W* the internally stored weight.

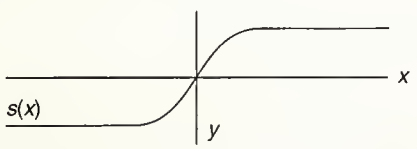


Figure B. Sigmoid function.

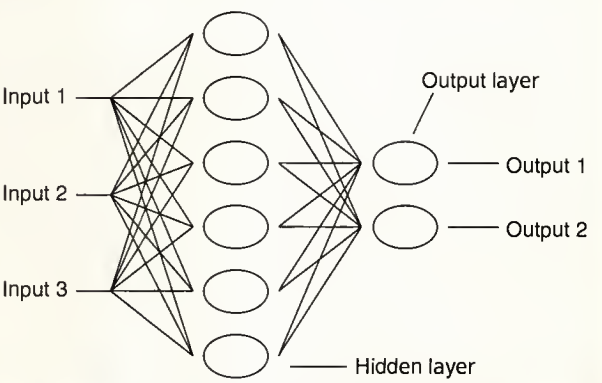


Figure C. Two-layer network.

For example, the physical interaction between two colliding objects can be quite complex and depend on the objects' composition, shape, velocity, and angle of incidence. A major advantage of using a neural network is that it is not necessary to fully understand the underlying physics of the system. However, this is only true if a representative set of examples can be generated. The neural network uses its associated learning algorithm to learn from examples.

The ideas we present can be applied to a broad class of real-world signal processing problems, and the techniques can be applied in applications that recognize, compress, decompress, or map time-domain waveforms into control variables.

Our impact recognition application is to observe a collision between two objects, one known and one variable, and immediately classify (or identify) the variable object based on the observations. In our case, the known object is made of wood, and the variable object is made of some other material. An accelerometer mounted on the known, wooden object observes the collision. Figure 2 shows the circuit supplying power to the accelerometer and conditioning its output signal.

Experimental set-up

The neural network that recognizes the objects has five layers, the first of which is a DFT module. The second and third layers make up the magnitude module, and the fourth and fifth layers are the recognition module. The three layers that make up the DFT and magnitude modules form the preprocessing neural network. This preprocessor transforms the input signal from a time domain to a frequency domain representation. The preprocessor outputs then become the inputs to the two-layer recognition module, which classifies the frequency domain pattern and identifies the object that caused it.

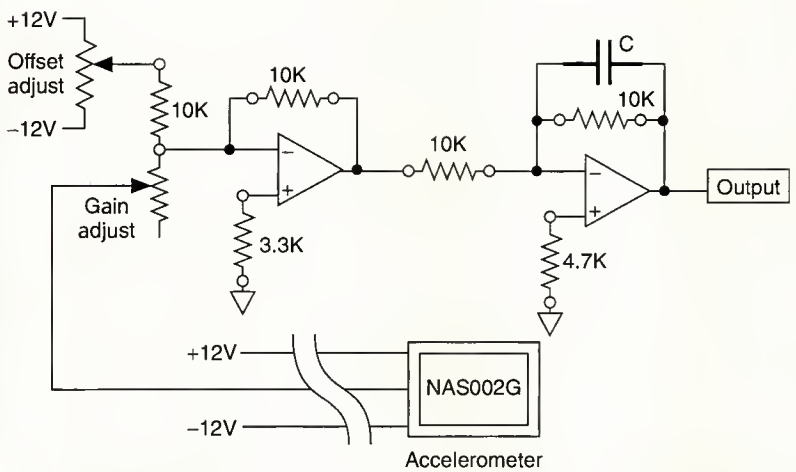


Figure 2. Accelerometer interface circuit. The first stage provides variable gain and offset. The second stage is an antialiasing filter.

Layer 1, the DFT

The discrete Fourier transform is defined as follows:

$$F(m) = \sum_{n=0}^{N-1} f(n) \cos \frac{2\pi mn}{N} + i \sum_{n=0}^{N-1} f(n) \sin \frac{2\pi mn}{N} \qquad m = [0 \ldots N-1] \tag{1}$$

$$f(n) = f(t)\Big|_{t=nT} \qquad n = [0 \ldots N-1]$$
$$F(m) = F(\omega)\Big|_{\omega=\frac{2\pi m}{NT}} \qquad m = [0 \ldots N-1] \tag{2}$$

where $f(n)$ is the nth sample of the input waveform, $F(m)$ is the mth component of the DFT, T is the time between samples, and N is the total number of samples taken.

Since neural networks map vectors, it is natural to use a neural network to implement the DFT, as shown in Figure 3a. In this figure circles represent the neurons. Each neuron's output is a function of the weighted sum of that neuron's inputs. A typical neural network produces a sigmoidal output; however a DFT network produces a linear output.

Figure 3b shows specifically how the 80170NX Electrically Trainable Analog Neural Network chip implements the DFT.[3] This representation emphasizes the matrix multiply operation that the net performs symbolically. Normally, the sigmoidal transfer characteristics of the 80170NX neurons are used; however, to permit an accurate approximation of the DFT, we make them as linear as possible by setting the voltage on the gain control pin as low as possible. See the box on p. 38 for more information on the 80170NX.

Most neural networks compute the weights by some sort of training algorithm. The DFT network calculates the correct weights ahead of time from the definition of the DFT. From Equation 1 and referring to Figure 3b, we achieve

$$W_{mn} = \cos 2\pi mn/N$$
$$V_{mn} = \sin 2\pi mn/N \tag{3}$$

where m equals $(0 \ldots N-1)$, and n equals $(0 \ldots N-1)$.

Once these values are loaded into the synapses of the 80170NX, the DFT network is complete. Note that since $F(m)$ is complex, the DFT network actually produces $2N$ outputs, N real and N imaginary (see Figure 3b).

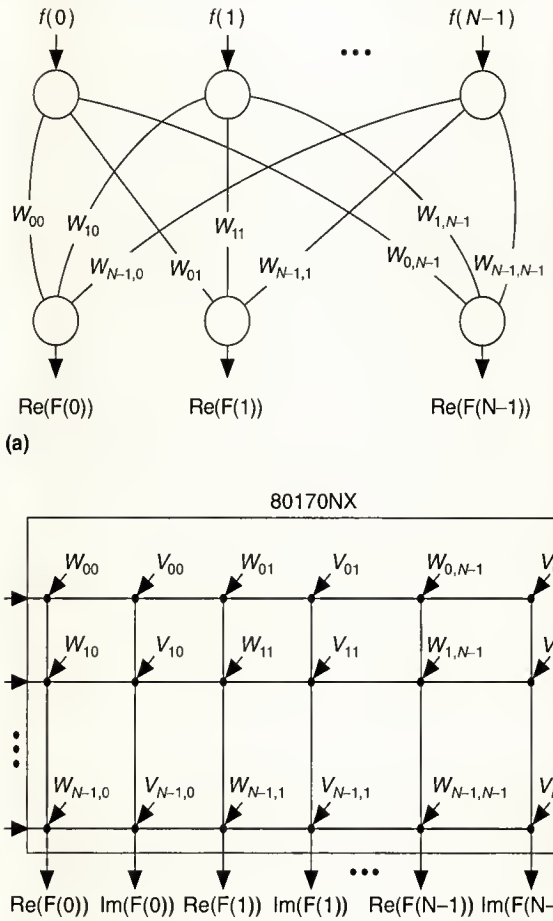


Figure 3. Implementing the DFT with a general neural network (note that the imaginary part is not shown) (a) and with the 80170NX (b).

Because of the symmetry properties of the DFT, however, only the first $N/2$ real and $N/2$ imaginary outputs contain nonredundant information.[4] We do not calculate redundant DFT outputs in this application.

Implementing the DFT network. First, the network creates a matrix that contains the desired DFT weights. The C code shown in Figure 4 generates this matrix. The code uses Equation 3 with $N = 32$ to calculate the weight values and stores them in a matrix format. A line containing the 64 biases (all set to zero) is included at the end of the file. DynaMind, a software package provided with the Intel Neural Network Training System (iNNTS),[5] loads the network into memory and then writes the weight matrix into the 80170NX. Weights are stored in analog EEPROM on chip and are programmed by applying 12V to 18V pulses.

Using the EMB. We implement this DFT network on the ETANN multichip board, EMB. The EMB is a development tool Intel provides to facilitate design of embedded neural network applications that may use up to eight 80170NX chips. EMB allows direct access to network inputs and outputs during prototyping while maintaining an umbilical connection to the iNNTS. The iNNTS manages pattern files and network interconnectivity, performs back-propagation learning, and controls weight setting. The EMB and all associated hardware used in the impact recognition system are shown in Figure 5 on p. 39.

Network reconfiguration. The inputs and outputs of the 80170NXs can be connected to buses on the EMB. These buses connect to the iNNTS adapter so that the host computer, using an addressing scheme, can read the outputs from any chip and write them to the inputs of any chip. This arrangement allows reconfiguration of the neural network implemented on the EMB under control of the neural network simulators, DynaMind and BrainMaker, or other user-written software running on the iNNTS host computer.

```
int             i, j, N = 32;
double          w, arg;
FILE            *outf;
        ⋮
for (i=0; i < N; i++)
        {
        for (j=0; j < N; j++)                   /* real */
                {                               /* part */
                arg = (2.0*PI*i*j /N);
                w = 2.5*cos(arg);
                fprintf(outf, "%.4f ",w);
                }
        fprintf(outf,"\n");

        for (j=0; j < N; j++)                   /*imag */
                {                               /* part */
                arg = (2.0*PI*i*j /N);
                w = -2.5*sin(arg);
                fprintf(outf,"%.4f ",w);
                }
        fprintf(outf,"\n");
        }
fprintf(outf,"\n");
for (j=0; j < 2*N; j++)                         /* biases */
        fprintf(outf,"%.4f ",0.0);
        ⋮
```

Figure 4. C code for generating a matrix with the DFT coefficients.

The 80170NX

Intel's 80170NX, which is also known as the Electrically Trainable Analog Neural Network (ETANN), offers developers of neural network applications an extremely fast and flexible analog architecture. Its on-board circuitry implements 64 neurons with 10,240 synapses in two arrays of 80 inputs, which are subdivided into 64 analog inputs and 16 analog bias inputs. Figure D illustrates the structure of one such neuron.

Inputs come into the 80170NX in the form of analog voltages. Each input can be thought of as a component of an input vector. Similarly, for each neuron the weights can also be thought of as components of a vector. Depending on the configuration in which the chip is used, the dimension of the vectors is either 64 or 128.

Weights are stored by the 80170NX as analog voltages on nonvolatile, electrically alterable cells. Each input to the chip is multiplied by its corresponding weight. The product is in the form of a current. For every one of the 64 neurons, currents from the multipliers are independently summed, and the result is then made nonlinear by a mapping to the sigmoid function. The sigmoid's gain is controlled either by setting a gain control voltage or by selecting the high gain mode that maximizes the sigmoid's slope in the transition region, effectively making the outputs digital.

The 80170NX is composed of two synapse arrays, both of which have hold circuitry on the inputs. One of the arrays, however, gives the 80170NX its flexibility in implementing a variety of networks. It can be configured as either a second on-chip network layer or used to double the number of inputs to each neuron to 128 (see Figure E). Because the 80170NX's inputs and outputs are fully compatible, multiple 80170NXs can be cascaded to produce more complex networks.

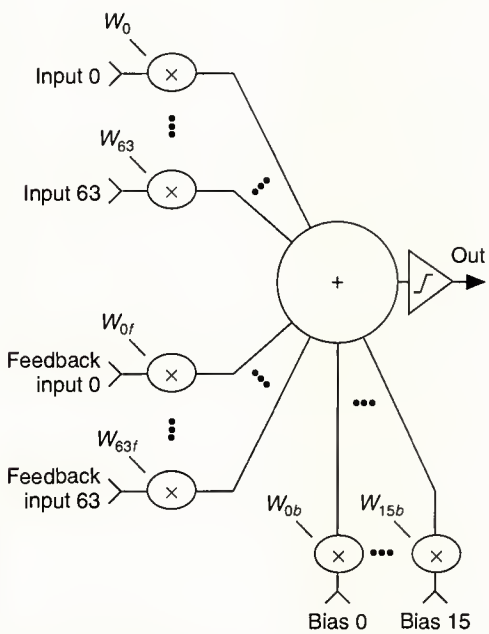


Figure D. The 80170NX neuron structure.

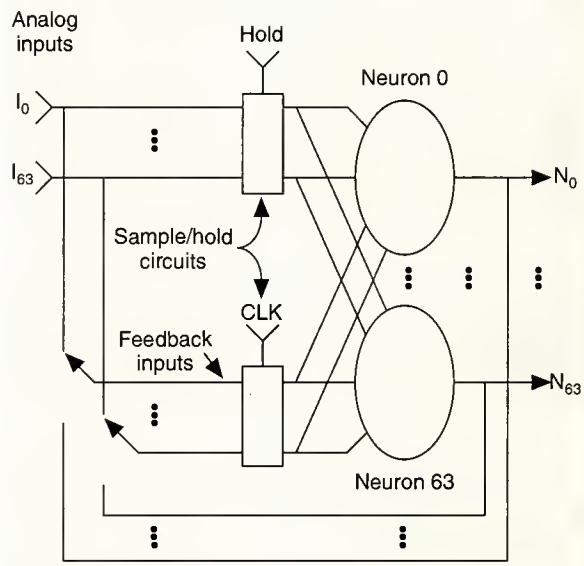


Figure E. Architecture of the 80170NX.

DFT network performance. The circuit shown in Figure 6 uses the Reticon RT0032AN analog tapped delay line[6] to produce the 32 inputs to the DFT network. The RT0032AN's sampling frequency may be varied by the frequency of the clock used to drive it.

Figures 7, 8, and 9 show two of the DFT network outputs, F(2) and F(3), when the input is a sinusoid of varying frequency. The sample rate of the delay circuit is set by SG1 (Figure 6) to 32 kHz. From the sample rate we can calculate the frequency resolution of the network outputs (on p. 40):[7]

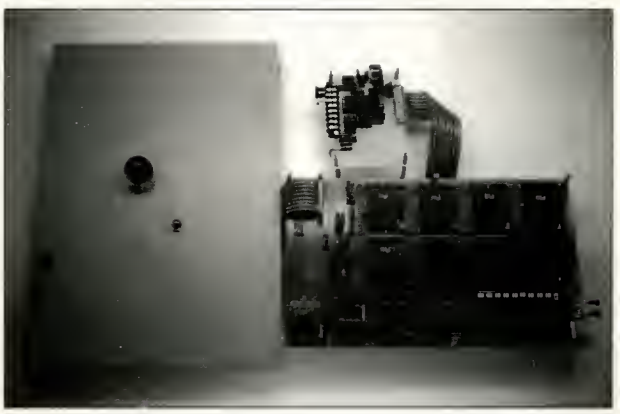


Figure 5. The impact recognition system. The EMB in the foreground contains five socketed 80170NX chips. The iNNTS, used only during training, is not shown.

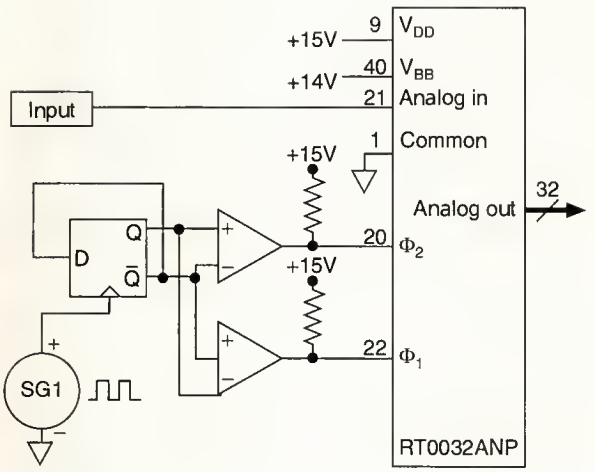


Figure 6. Analog tapped delay line. The RT0032ANP samples the input at a rate set by SG1. Its outputs are the 32 most recent samples. Note that the outputs are analog.

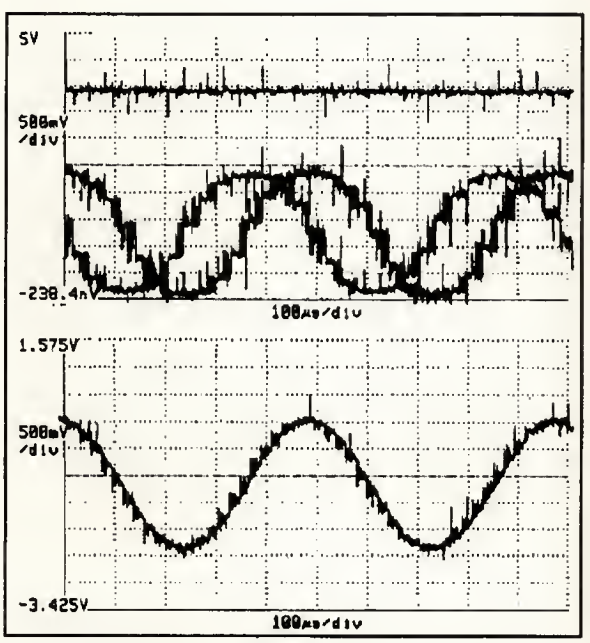


Figure 7. DFT network response to a 2.0-kHz sinusoid. The bottom trace is the input to the tapped delay line; the middle traces are neurons 4 and 5, Re[F(2)] and Im[F(2)]; and the top trace is neuron 6, Re[F(3)].

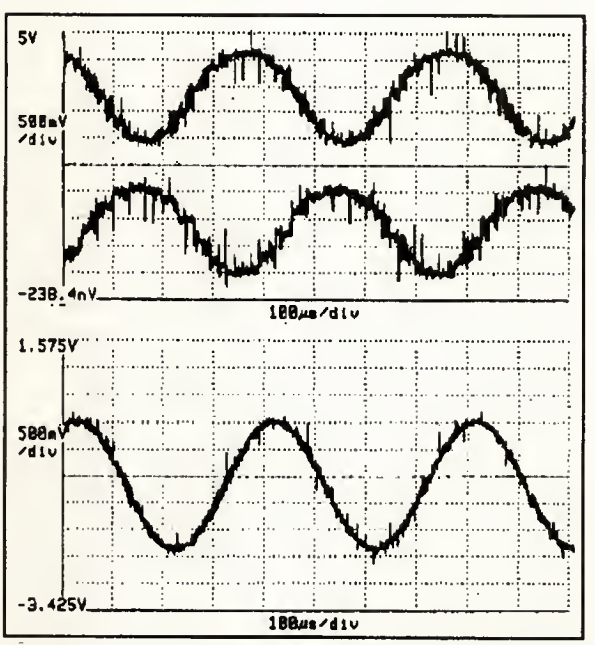


Figure 8. DFT network response to a 2.5-kHz sinusoid. The bottom trace is the input to the tapped delay line; the middle trace is neuron 4, Re[F(2)]; and the top trace is neuron 6, Re[F(3)].

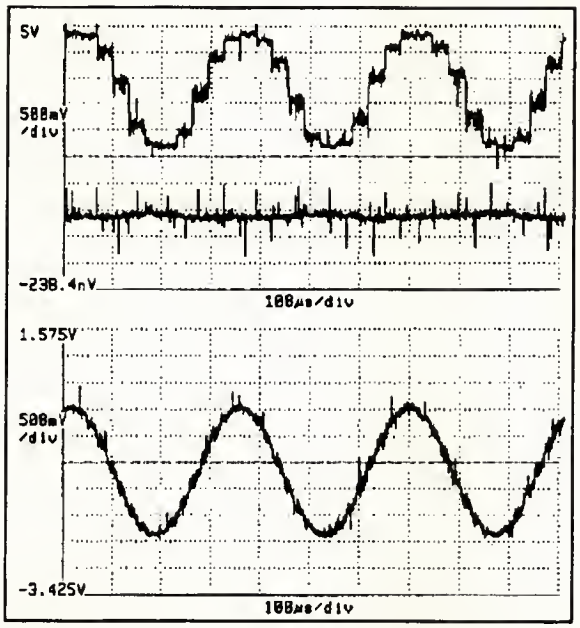


Figure 9. DFT network response to a 3.0-kHz sinusoid. The bottom trace is the input to the tapped delay line, the middle trace is neuron 4, Re[F(2)]; and the top trace is neuron 6, Re[F(3)].

$$f_{res} = f_{samp}/N = 32 \text{ kHz}/32 = 1 \text{ kHz} \quad (4)$$

In Figures 7, 8, and 9, the bottom trace is the input waveform, and the middle and top traces are Re[F(2)] and Re[F(3)]. Figure 7 also shows Im[F(2)].

With an input frequency of 2.0 kHz (Figure 7), F(2) becomes active, F(3) is zero, and Re[F(2)] and Im[F(2)] are 90 degrees out of phase. When the input frequency is 2.5 kHz (Figure 8), both F(2) and F(3) are partially active. When the input frequency is 3.0 kHz (Figure 9), only F(3) is active.

Layers 2 and 3, generating a magnitude spectrum

A measure of the total signal present at each frequency is necessary to perform frequency domain-based recognition. To calculate this, we combine the information contained in the real and imaginary parts of the DFT at each frequency. The magnitude module we describe performs this task. $|F|$ designates the output of the magnitude module we call the magnitude spectrum.

Learning the magnitude computation. We used a set of 77 patterns to train a network to compute the magnitude of a two-dimensional vector, as displayed in Equation 5:

$$N_0 = \sqrt{(I_0^2 + I_1^2)} \quad (5)$$

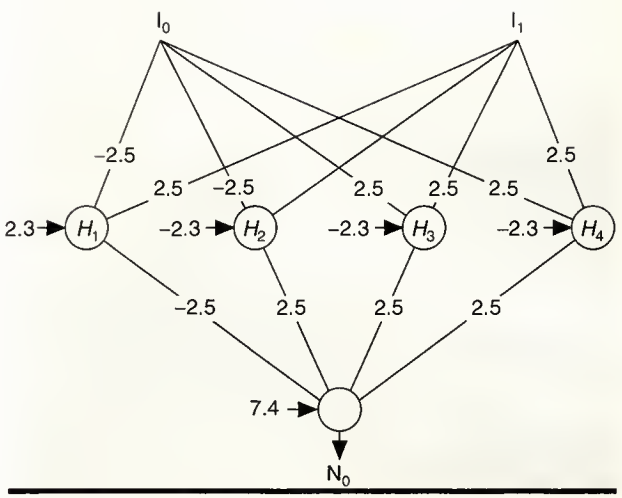


Figure 10. Neural network implementation of the magnitude calculation shown in Equation 5. (*H* = hidden layer.)

Each training pattern has two inputs and one target output. The inputs range from −1.0 to +1.0 and cover the region bounded by the unit circle with roughly equally spaced points. The target outputs are scaled and offset to use the full range of the 80170NX's outputs (0V to 3V). Thus, an output of 0V corresponds to a magnitude of 0, and an output of 3V corresponds to a magnitude of 1. The architecture of the magnitude network shown in Figure 10 consists of one hidden layer with four neurons. Note the presence of bias inputs to each neuron. These are an important part of the magnitude network.

We used the Madaline III algorithm[8] to train the network. After several seconds of training in simulation, the network converges to a solution that does a good job of estimating the magnitude of the two inputs. There are several equivalent solutions to this problem; Figure 10 shows the weight set we used.

Some insight into the operation of the magnitude network can be gained by carefully examining Figure 10. Each neuron in the hidden layer performs a type of selective communication with the output neuron. The third and fourth neurons, for example, map the sum of I_0 and I_1 as follows:

$$\begin{aligned} H_3&: \quad (-1 \ldots 0) \rightarrow -1, (0 \ldots +1) \rightarrow (-1 \ldots 0) \\ H_4&: \quad (0 \ldots +1) \rightarrow (0 \ldots -1), (0 \ldots +1) \rightarrow -1 \end{aligned} \quad (6)$$

Note that if we now add the value +1 to the outputs of H_3 and H_4, and sum the outputs, we have created a subnetwork that computes the absolute value of $I_0 + I_1$. This is effectively what the output neuron does.

Hidden-layer neurons 1 and 2 work in the same manner, only they operate on the difference between the inputs ($I_0 - I_1$)

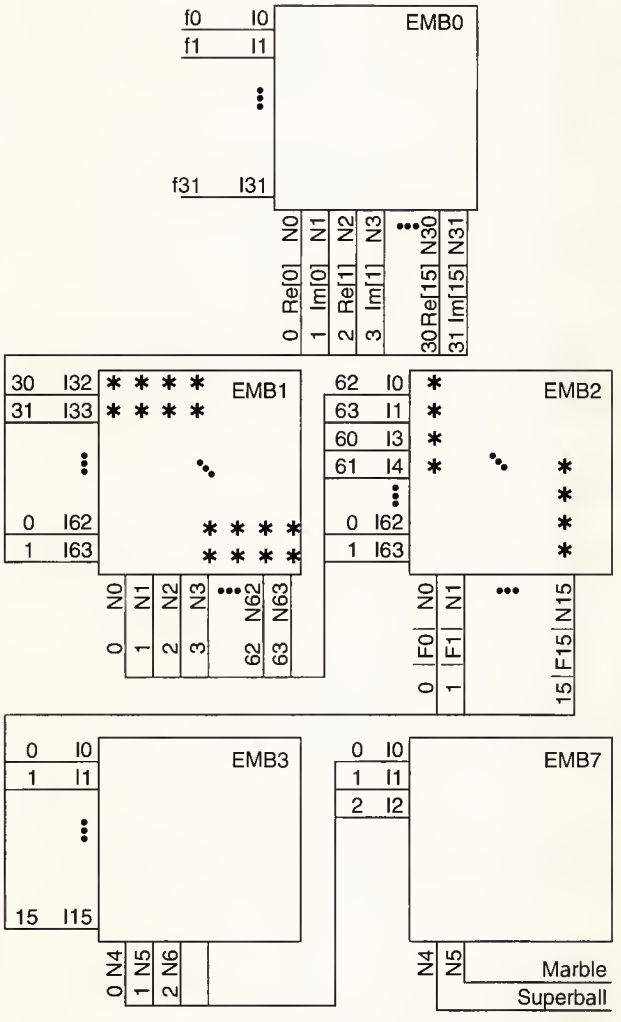


Figure 11. Schematic showing connections between the sockets on the EMB, which houses the various parts of the five-layer network.

instead of the sum. All of the combined information at the output neuron forms an estimate of the magnitude of the inputs.

Magnitude network into hardware. To create the full magnitude spectrum network, the two-input, one-output magnitude network must be copied 16 times and written to hardware. One 80170NX can implement networks with multiple layers but to do so requires a simple state machine to control the internal clocking of data.[3] To avoid this extra hardware, we implement the two-layer magnitude network on two 80170NX chips, one layer per chip.

As is shown in Figure 11, inputs I_{62} and I_{63} of EMB socket 1 drive neurons N_{60} through N_{63} (as indicated by the stars in Figure 11). I_{62} and I_{63} are the two inputs to the first of the 16 magnitude minimodules. Neurons N_{60} through N_{63} are the four hidden-layer neurons of this minimodule, and they connect to EMB socket 2, inputs I_0 through I_3. These inputs drive neuron 0, which is the output neuron of the first magnitude minimodule designated $|F0|$ in Figure 11. The remaining 15 magnitude minimodules are arranged in a block-diagonal manner.

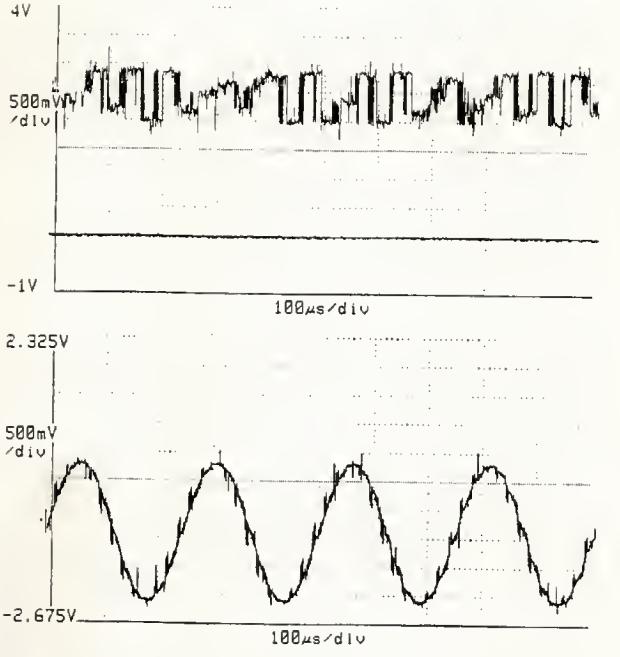


Figure 12. Magnitude network response to a 4.0-kHz sinusoid. The bottom trace is the input; the middle and top traces are neuron 3 $|F(3)|$ and neuron 4 $|F(4)|$.

Once a module has been downloaded, we train it in a chip-in-loop (CIL) style to adjust the weights to compensate for variations in the analog computing hardware of the 80170NX. We set the learning rate very low to assure the solution does not diverge. This procedure repeats until all 16 magnitude minimodules have been downloaded and CIL trained.

The three-layer preprocessing network is now complete. A 4-kHz sinusoid is input to the tapped delay line. Figure 12 shows two of the outputs of the preprocessing network. Note that although $|F(4)|$ does not stay constant at +3V, it is easy to see that N_4 detects significant energy at 4 kHz while N_3 detects little energy at 3 kHz.

Layers 4 and 5, object recognition

Once the spectral energy density patterns have sufficient information to allow discrimination of different objects, we can classify the objects with layers 4 and 5 of the network.

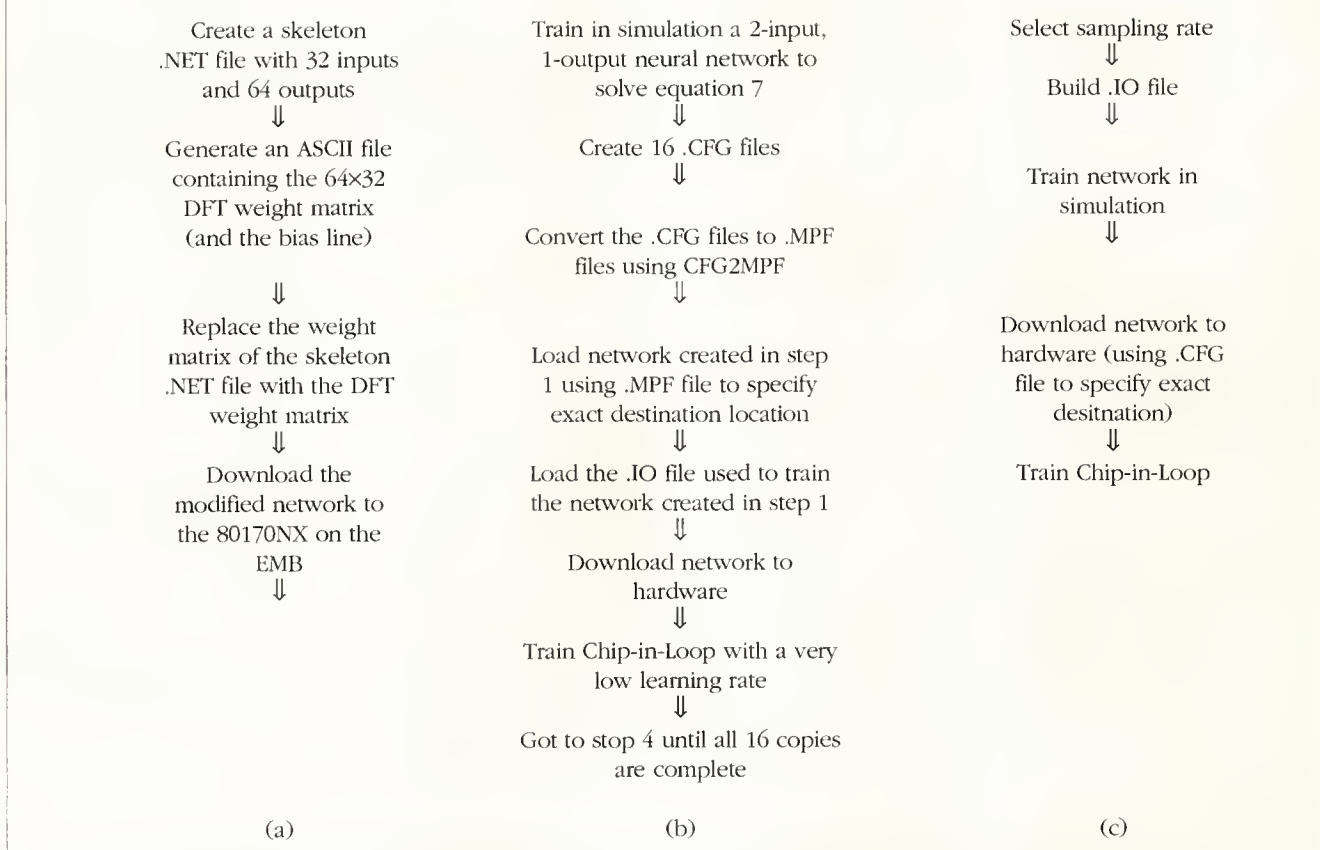


Figure 13. Flow chart showing design procedure of the DFT (a), the magnitude (b), and the recognition modules (c).

Table 1. System response.

Output 0	Output 1	Event
0	0	Nonevent
0	1	Marble
1	0	Superball

Figure 13c summarizes the procedure for training a neural network to perform this task. (Figure 13a,b describes the design procedures of the DFT and magnitude modules.) The first step is to choose a sampling rate. At this point, the network is independent of this value, but the remaining layers must be trained to a specific sampling rate. A value of 11.2 kHz is high enough to prevent aliasing of the highest modes caused by the objects but low enough to provide adequate frequency resolution. We calculate the frequencies corresponding to the network outputs by inserting 11.2 kHz into Equation 4 and then calculate the spacing of the spectral bands and multiply this by the index m.

$$\begin{aligned} m &= 0 \ \mathrm{F}(0) \rightarrow 0 \ \mathrm{Hz} \\ m &= 1 \ \mathrm{F}(1) \rightarrow 350 \ \mathrm{Hz} \\ m &= 2 \ \mathrm{F}(2) \rightarrow 700 \ \mathrm{Hz} \\ &\vdots \\ m &= 15 \ \mathrm{F}(15) \rightarrow 5{,}250 \ \mathrm{Hz} \end{aligned} \tag{7}$$

Next we build a training set. The training inputs for the recognition network are the outputs of the magnitude spectrum network. The training outputs are two numbers whose values correspond to the object that was dropped. (See Table 1.)

We also included some nonevent data to ensure that neither output will turn on when no object was dropped. We used a storage oscilloscope to look at the outputs of the magnitude spectrum network when the different objects are dropped. Note that outputs 1, 2, and 5, corresponding to 350, 700, and 1,750 Hz, are the most active. This resulted in training patterns with three inputs and two outputs. We read the

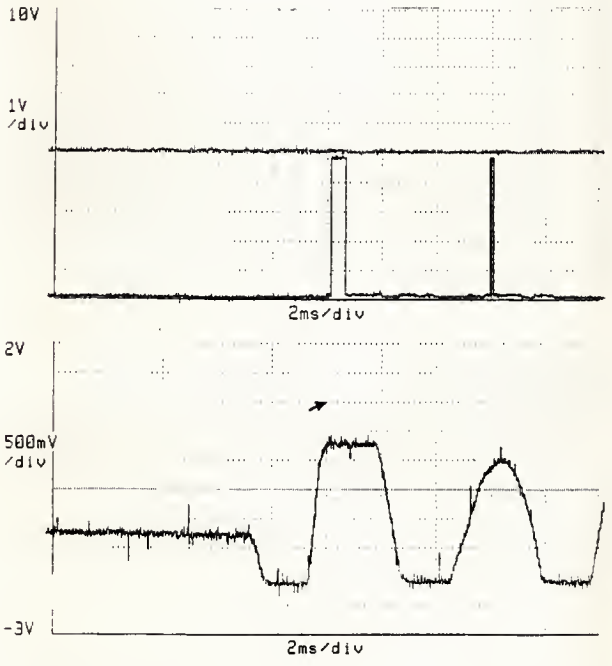


Figure 14. Recognition network response to the superball. The bottom trace is the accelerometer output. The middle and top traces are neurons 4 and 5 (superball and marble) in Figure 11.

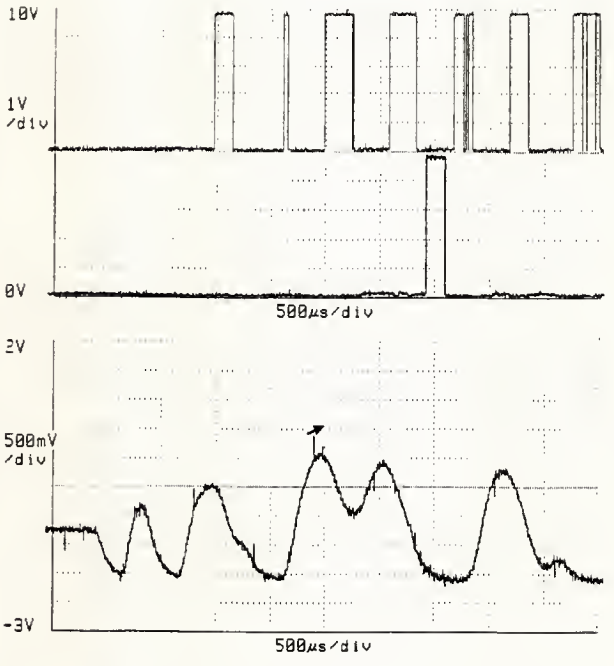


Figure 15. Recognition network response to the marble. The bottom trace is the accelerometer output. The middle and top traces are neurons 4 and 5 (superball and marble) in Figure 11.

values of the three inputs on the storage oscilloscope as the different objects repeatedly dropped.

We are working to sample and hold the entire 16-component magnitude pattern and read it using an A/D converter in the iNNTS to rapidly capture many patterns. With automated gathering of training data, we can quickly implement the recognition of other objects, capture patterns, and retrain the network.

We've placed most of the data collected in a file for training purposes (both simulation and CIL) and reserved about 10 percent of the data to test the network once its performance on the training set is acceptable.

The architecture of the recognition network is now almost completely defined. With three inputs, two outputs, and the assumption that one hidden layer will be enough, all that remains is to decide on the number of neurons in the hidden layer. Using DynaMind, we find that a two-layer network with three hidden-layer neurons will adequately perform the recognition task. As in the magnitude module, we will realize this network on two 80170NX chips, one for the hidden layer (housed in EMB socket 3) and one for the output layer (housed in EMB socket 7). The previous Figure 11 showed schematically the wire-wrap connections made on the EMB.

Note that although only outputs N1, N2, and N3 of the magnitude spectrum network are used in the classification; all 16 outputs connect to the fourth layer. Thus, if the set of objects to be classified changes in any way, modifications need only be made to the weights of the last two layers of the network.

Figures 14 and 15 show the results of the five-layer object recognition network. Note that it takes less than 3 ms to recognize the superball. This is impressive since the time it takes to fill the 32 outputs of the delay line is 2.86 ms. The network recognizes the marble in about 1 ms. The features of the DFT that characterize the marble apparently become prominent even before the delay line has completely filled with data.

Once an object has recoiled from the platform with the accelerometer attached, the frequency of the vibration measured by the accelerometer decreases toward a resonant frequency of the platform. Since the neural network is trained only on the data generated during impact, its output after impact is unpredictable. This accounts for the pulses that occur in Figures 14 and 15 after the first pulse. Latches capture the first pulse and "lock out" subsequent spurious pulses related to the resonances in the accelerometer platform.

The delay of the 80170NX is 3 µs per layer, independent of

how many connections are used on a chip. Thus, for a five-layer network, we expect a processing time of 15 µs. The processing speed of this network is more than 50 times that required by this application.

OUR HYBRID SYSTEM IDENTIFIES objects based on the vibrations caused when objects impact a platform with an accelerometer attached. The system uses a conventional DFT and a multilayer Perceptron, both of which are implemented using a total of five 80170NX devices residing on a multichip prototyping board.

We've divided the five layers of processing that recognize the objects into a DFT module, a magnitude module, and a recognition module. This structure reduced the amount of effort needed to collect data and train the recognition network. It also allowed the preprocessing neural network formed by the DFT and magnitude modules to be readily reused for other applications by just capturing new patterns and retraining.

The parallel architecture of the 80170NX produces a delay of only 15 µs for the five layers of processing. This performance and our design approach provide a solution that is fast enough and flexible enough to solve a wide range of real-time signal processing problems.

Neural networks are well on their way to becoming a standard tool for signal processing. Their use is growing because their nonlinear characteristics allow them to provide better solutions such as reduced error rates in hand-printed character recognition systems.[9] They are also finding favor because the learning algorithms used to determine the weights for a neural network can save engineering time if a good set of example data is available.

Neural networks have been implemented using DSPs such as the Intel 860[10] and the Texas Instruments TMS320C30[11] as well as specialized neural network hardware such as the 80170NX and Adaptive Solutions' CNAPS architecture.[12] Other general-purpose neural network chips such as Siemen's MA16 chip[13] and the Intel-Nestor chip funded by the US Defense Dept.'s Advanced Research Projects Agency (DARPA)[14] are also likely to become available. However, in the near term the most likely embodiment of neural networks that will find use in high-volume commercial applications is in the form of function or application-specific ICs (FASICs). Unfortunately because of the competitive advantage that neural network technology can provide and the confidential nature of most FASIC designs, the most successful near-term applications of neural networks are likely to remain concealed. For example, the implementation of a neural network in a cellular telephone to recognize spoken digits would likely not be reported.

Another reason for neural networks seeing early adoption in FASIC form is that the neural network is typically not the largest component of a system; pre- and postprocessing as well as other functions require as much or more computing power. If a FASIC is already being used, it will be more cost effective to integrate the neural network on the FASIC. As we've described, the similarity of neural network computations to linear signal processing functions should make integrating a neural network relatively easy, whether on the same chip with preprocessing or separately.

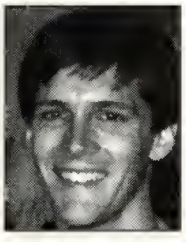
Acknowledgments

We thank Alan Baldwin and Gerhard Parker for their support and encouragement. We also acknowledge that Steven Deiss at Applied Neurodynamics designed and fabricated the EMB.

References

1. *Proc. IEEE Workshop Neural Networks for Signal Processing*, B.H. Juang, S.Y. Kung, and C.A. Kamm, eds., IEEE Press, Piscataway, N.J., 1991.
2. D.E. Rumelhart and J.L. McClelland, "Parallel Distributed Processing," *Explorations in the Microstructure of Cognition*, Vol. 1-3, MIT Press, Cambridge, Mass., 1986, pp. 3-110.
3. "80170NX Electrically Trainable Analog Neural Network Data Sheet," Intel Corp., Santa Clara, Calif., 1991.
4. L.R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Ch. 2, Prentice-Hall, Englewood Cliffs, N. J., 1975.
5. *iNNTS Neural Network Training System Users Guide*, Intel Corporation, Santa Clara, Calif., 1991.
6. "RT0032AN Tapped Analog Delay Data Sheet," EG&G Reticon, Sunnyvale, Calif., 1987.
7. R.N. Bracewell, *The Fourier Transform and Its Applications*, Ch. 10 and 18, McGraw-Hill, New York, 1986.
8. D. Andes et al., "MR3: A Robust Algorithm for Training Analog Neural Networks," *Proc. Int'l J. Conf. Neural Networks*, IEEE Press, 1990, pp. 533-536.
9. Y. Le Cun et al., "Handwritten Digit Recognition with a Back-Propagation Network," *Neural Information Processing Systems*, Vol. 2, Morgan-Kauffman, San Mateo, Calif., 1990, pp. 396-404.
10. *HNC Balboa 860 Card User's Manual*, HNC Inc., San Diego, Calif.
11. *Neural Emulation Tool User's Manual*, Loral Space Information Systems, Houston.
12. M. Griffin et al., "An 11 Million Transistor Digital Neural Network Execution Engine," *Proc. IEEE Int'l Solid-State Circuits Conf.*,

1991, pp. 180-181.
13. J. Ramacher et al., "Design of a First Generation Neurocomputer," *VLSI Design of Neural Networks,* Kluwer Academic, Boston, Mass., 1991, pp. 271-310.
14. M. Holler et al., "A High Performance Adaptive Classifier Using Radial Basis Functions," *Proc. Government Microcircuit Applications Conf.,* US Army Research Laboratory, Fort Monmouth, N.J., *1992, pp. 261-264.*



Jeff Brauch is a hardware development engineer at Hewlett-Packard in the Medical Imaging Systems Research and Development Laboratory. When this article was written, he worked in Intel Corporation's Neural Network Group. He has also worked as a staff member for the Massachusetts Institute of Technology Lincoln Laboratory. His technical interests include neural networks, digital signal processing, and automatic control systems.

Brauch received his BS degree from the University of Colorado and his MS from the University of Minnesota, both in electrical engineering.

Simon M. Tam is a senior architect in Intel's Neural Network Group, where he has developed EEPROM and flash memory technologies. He coinvented the ETOX flash memory cell and codeveloped the ETANN neural network chip. His current interests include neural network implementations and speech and pattern recognition using neural networks.

Tam received the BS, MS, and PhD degrees in electrical engineering and computer science from the University of California, Berkeley. He has authored or coauthored over 20 technical articles and been awarded 10 US patents. He is a member of Tau Beta Pi and the IEEE.

Mark A. Holler is technical director of Intel's Neural Network Group, where he investigates new applications for analog VLSI neural network hardware and design of their components. His interests include developing neural architectures for elastic matching tasks such as speech and handwriting recognition.

Holler received his BS and MS degrees in electrical engineering from the University of Wisconsin. He is a member of the IEEE, has been granted 15 patents, and has authored or coauthored 25 papers in the areas of EPROMs and neural networks.

Arthur L. Shmurun is an engineer in Intel Corporation's Neural Network Group. His technical interests include VLSI systems and applications of neural networks to signal processing and pattern and speech recognition. Shmurun received his BS degree from the California Institute of Technology in Pasadena, California.

Direct questions concerning this article to Arthur L. Shmurun, Intel Corporation, Neural Network Group, RN-317, 2200 Mission College Blvd., Santa Clara, CA 95052-8119; or e-mail at ashmurun@sc9.intel.com.

Figure-Ground Segregation Using an Analog VLSI Chip

Our working, analog VLSI vision chip labels all points inside a given contour with one voltage and all remaining points outside this contour with another voltage. Its behavior is very robust, since small breaks in the contour are automatically "sealed," providing for figure-ground segregation in a noisy environment. This circuit with its networks of resistors and switches represents a step toward object-level processing, since a single voltage value encodes the property of an ensemble of pixels.

Jin Luo

Tanner Research Inc.

Christof Koch

California Institute of Technology

Bimal Mathur

Rockwell International Science Center

Human observers effortlessly perform figure-ground segregation. They can determine whether a specified point in their visual field is inside or outside of one (or more) closed visual contours. Such an object-based vision algorithm could be implemented by means of a conventional numerical analysis method running on digital computers. However, this method is not feasible for real-time applications. As an alternative solution to this typical vision problem, we describe an analog, parallel, computational system built on a single, power-lean, CMOS VLSI chip that labels all points inside a possibly incomplete and noisy contour in real time.

Background

Early vision consists of the set of processes that recover physical properties of a visible three-dimensional surface, such as its distance from the observer or its surface texture, from two-dimensional intensity data. The associated algorithms are typically based on pixels, in the sense that a scalar or vector is computed at each picture element in the scene (for instance, edge detection and optical flow algorithms produce an output at every grid point).

Over the last several years, we and others have successfully designed and built a number of analog CMOS (complementary metal oxide semiconductor) VLSI circuits with on-chip photoreceptor arrays that implement such pixel-based algorithms.[1-3] Here we discuss one instance of a new class of circuits that outputs a single variable associated with a contour or an entire object in the image.

Horn[4] first raised the idea of using analog, nonclocked circuits for solving vision problems. Horn proposed the use of a hexagonal grid of resistances to find the inverse of the discrete approximation to the Laplace transform. Poggio and Koch[5] discussed a group of image processing algorithms known as standard regularization algorithms that map onto simple resistive networks. (Most early vision algorithms can be cast in this form. The "optimal" solution can be found by minimizing a cost function incorporating various generic constraints, such as "surfaces should be piecewise smooth."[6,7])

Exploiting Kirchhoff's and Ohm's law, Poggio and Koch proved that the minimum of the regularized, quadratic cost functional is equivalent to the state of least power dissipation in an appropriate linear resistive network. Here, injectors connected to certain nodes represent the data, and the steady-state voltage distribution provides the solution. In other words, for each such quadratic cost functional an associated resistive network exists, whose steady-state voltage distribution corresponds to the minimum of the

0272-1732/92/1200-046$03.00 © 1992 IEEE

MOS transistor's subthreshold operation

A number of circuits for image processing, particularly those based upon Mead's subcircuit types and design practices,[10] use an ordinary CMOS process with the transistors running in the subthreshold range. Figure A1 shows the drain current I_d as a function of source-drain voltage for four different values of the gate-source voltage. After I_d increases rapidly with drain voltage, it saturates to a nearly constant value—independently of drain voltage. The slight slope in the saturation region is caused by the change of effective channel length with drain voltage.

The saturation current in Figure A1 is plotted in Figure A2 as a function of gate-source voltage, where drain voltage is fixed at 2V. The current is an exponential function of gate voltage over five or more orders of magnitude. At about 0.9V the threshold voltage is reached, after which value the drain current becomes a quadratic function of the gate-source voltage. This exponential nonlinearity in the subthreshold voltage regime is ideal for building a variety of computational primitives.[10] Another advantage of operating in the subthreshold range is its very low power dissipation (less than 100 mW for a typical 48×48 vision-processing network chip).

With these circuit elements as building blocks, we have designed a large number of successful vision chips.

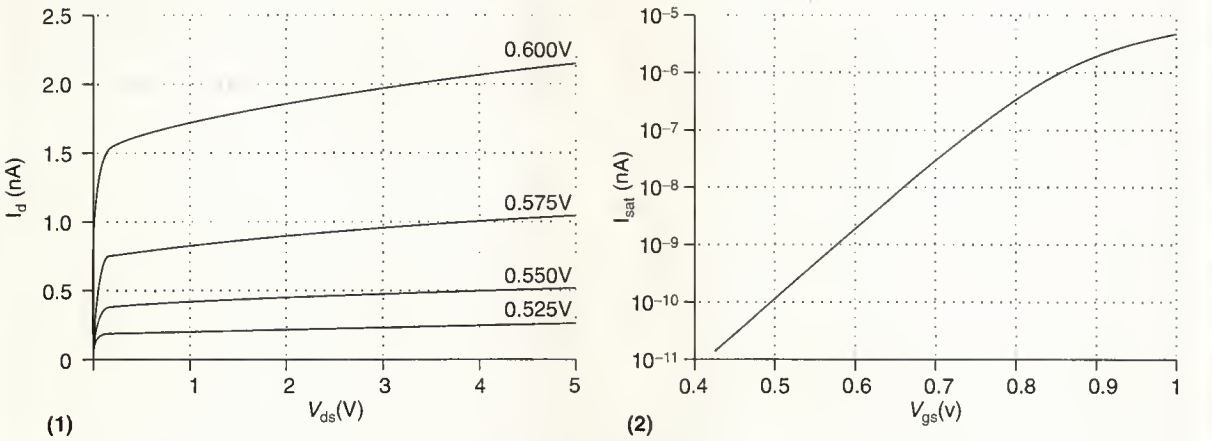


Figure A. Measured I-V characteristics of a MOS transistor operating in the subthreshold region.

variational functional. Thus, instead of programming a powerful and completely general-purpose von Neumann machine (as in a digital computer), the physics of resistive networks derive a solution to the early vision problem. These circuits have been generalized to include nonlinear circuit elements, where the steady-state voltage distribution corresponds to minimizing a nonconvex variational functional (for example, see Harris et al.[8,9]).

The development of subthreshold, analog VLSI circuits for various sensory tasks by Mead[10] (see above box) enabled us to implement resistive networks for solving early vision problems. Two circuit elements are particularly attractive for image processing. A photo-transistor with a logarithmic voltage outputs over five orders of light intensity. That is, a photoreceptor converts the incoming irradiance into a voltage value using a logarithmic mapping[10,11] and a nonlinear resistor called the horizontal resistor (HRes). This is a small transistor circuit with a quasilinear current-voltage relationship.[10,12]

This nonlinear resistor circuit (highlighted in the HRes box, next page) implements a saturating resistance. The slope of the IV curve around the origin (that is, the effective resistance of the device) can be varied over several orders of magnitude. A large number of application-specific integrated circuits (ASICs) have been built out of the combination of these two circuit elements. Examples of such smart-vision circuits include chips for finding edges and for smoothing noisy data, for estimating motion and depth, and for locating outliers.[1] The chips usually include 1D or 2D arrays of photoreceptors or other analog input mechanisms and an array of resistive elements. These chips can be fabricated through the US DARPA-sponsored MOS foundry service, MOSIS.

Little work, however, has been done in building special-

The HRes horizontal resistor

Two key components of several image processing circuits are photoreceptors (not used by our circuit)[10,11] and resistances. Instead of laying down a resistive layer (using polysilicon or wells) to form a resistance with a fixed value, Mead[10] designed a circuit with less than 10 transistors that approximates a current-voltage relationship reminiscent of a resistance. This HRes circuit was first developed as a model of the horizontal cells in the retina and has two advantages over a passive resistor. The current saturates for large enough voltage differences, limiting the effect any one such circuit element can have on its neighbors. Also, in the linear region the value of the resistance can be varied over five orders of magnitude, from several hundred Kohms to 10 or so Gohms.

Figure B1 displays the circuit diagram of the resistive connection of HRes. Two pass transistors in series (T_1 and T_2) form an ohmic path between the two nodes. The schematic in Figure B2 depicts the bias circuit for HRes. This bias circuit corresponds to the two voltage sources between V_1 and V_{g1} and between V_2 and V_{g2}. The input V (corresponding to either V_1 or V_2) senses the voltage at one end of the resistive link. It generates an output V_g to bias the associated pass transistor in Figure B1. The bias circuit is an ordinary transconductance amplifier (see the Two Key box) connected as a follower, with an additional diode-connected transistor T_d. The output voltage V_g will follow V but with an offset equal to the voltage across T_d. In our resistive network designs, four pass transistors from the neighboring nodes share one bias circuit. Each HRes on the network is biased globally so that the network operates with a global space constant.

Figure B3 displays the simulated current-voltage characteristics of an HRes element. The current through HRes is linear for small values of the voltage gradient V_1-V_2, subsequently saturating for larger values. This I-V curve can be well approximated by a function of the form $I \propto I_{sat} \tanh [(V_1 - V_2)/2]$.

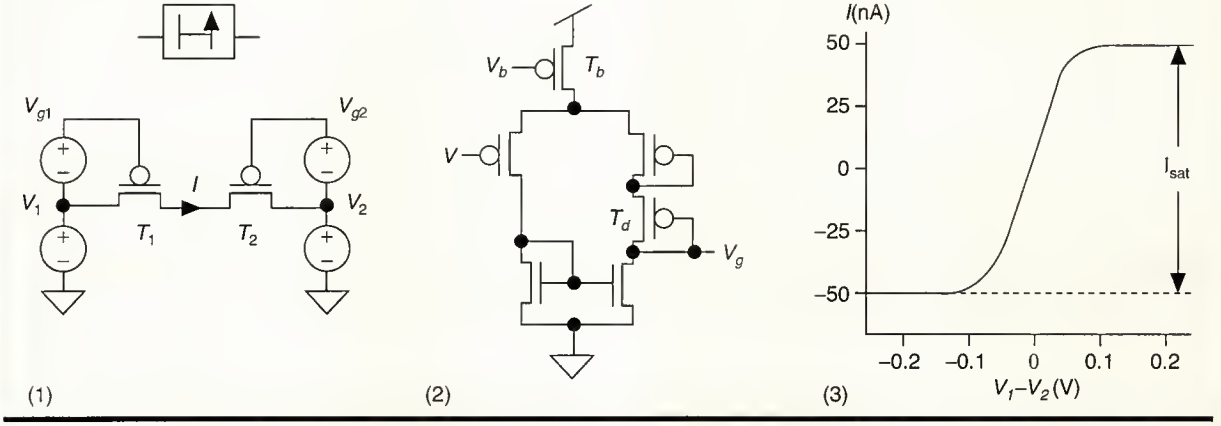


Figure B. Schematic diagram of a saturation resistor (1), its bias circuit (2), and the associated I-V curve (3).

purpose chips that move beyond these early-vision algorithms. The literature describes several chips that can be considered to be precursors to object-oriented chips. One of these computes the center of mass as well as the orientation of objects against dark backgrounds.[13,14] Another instance of an object-based analog vision chip is the Dynamic Wire circuit introduced by Liu and Harris.[15] This circuit estimates the total length of an unbroken contour supplied to a 2D resistive array. All of these produce a few outputs by integrating information from the entire image. Our circuit performs figure-ground segregation of a scene, labeling all the points inside a designated figure by one voltage and all other pixels outside this object using a different voltage value.

Figure-ground segregation

We were motivated to build this chip by the psychophysical observation that human observers effortlessly perform figure-ground segregation. That is, if shown a scene in which a small spatial region is distinguishable from the background by any number of visual features, such as brightness, depth, texture, or motion, humans rapidly label this region the "object" and everything else "ground," with little dependency on the length or the complexity of the outline of the object's contour. Ullman[16] argued in a seminal paper that this opera-

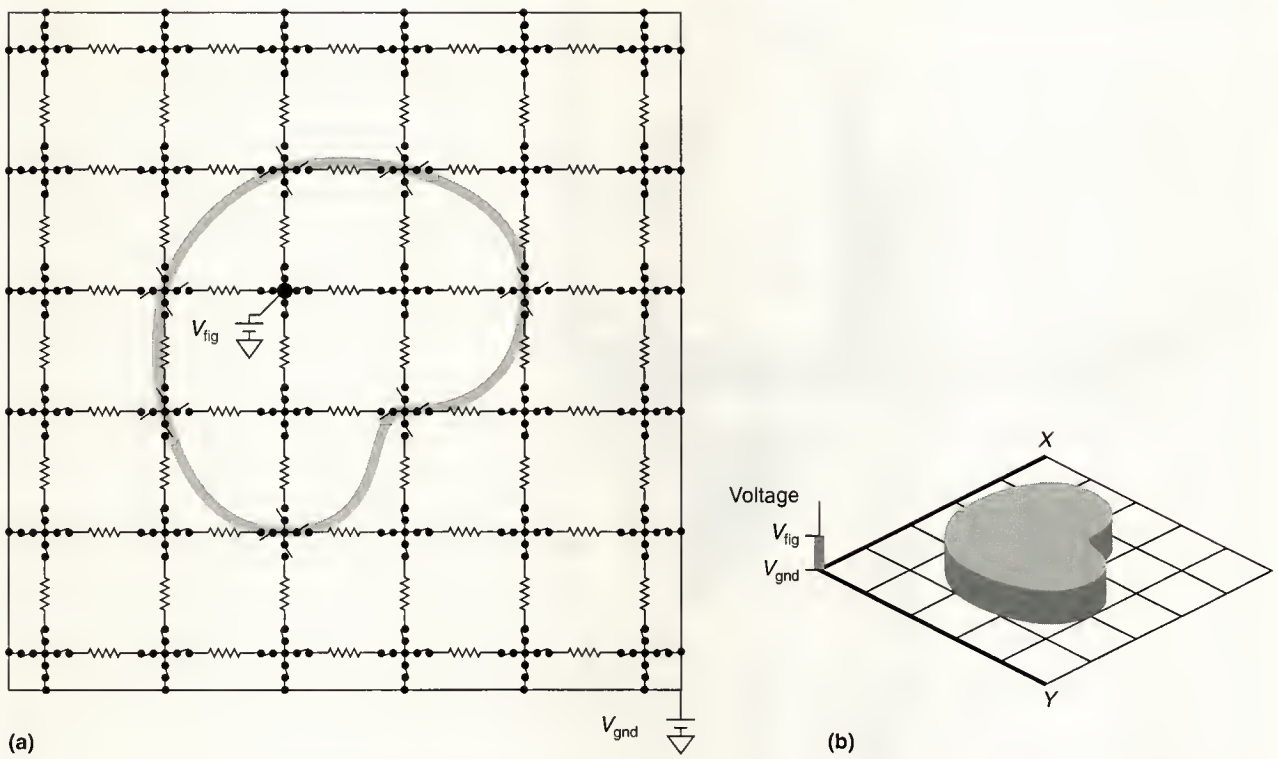


Figure 1. Schematic of the Figure-Ground resistive-grid chip: its resistors and switches (a) and a conceptual view of its segregation method (b).

tion as well as a number of related abilities of the human visual system constitute an elementary set of visual routines carried out by small modules within our visual cortex. Ullman's other "visual routines" include the shifting of the processing focus, indexing to an odd-man-out location, boundary tracing, and marking. At the time, Koch discussed with Ullman possible implementations of some of these operations using resistive networks.

We describe such a chip, which labels all points inside a given—possibly incomplete and broken—contour. Note that while the inspiration to build this chip was originally derived from biology, the resulting circuit bears no resemblance to any structure in the mammalian visual system.

The input to our Figure-Ground chip consists of a binary edge map, signaling the presence or absence of edges in the image. However, the current version of our chip does not include circuitry to capture the image or to compute the position of edges. This could be carried out using, for instance, a 2D version of the analog zero-crossing chip with the on-chip photoreceptors described earlier.[17] The binary output of such a chip, signaling the presence of a strong edge, would be scanned onto our Figure-Ground chip, where it would cause switches at the corresponding grid point within a rectangular resistive network to open. (See Figure 1.)

As shown in Figure 1a, the Figure-Ground network is made up of resistors and switches. At every grid point in the rectangular array where edges have been found, four switches are opened, isolating that node from its four neighbors (the shaded-edge contour corresponds to a series of isolated nodes). For our initial prototype chip, we assumed the visual contour will always encompass the center of the array. In other words, the figure to be segregated from the ground must enclose the central pixel of the circuit. At this center point, the resistive grid is connected to the battery V_{fig}, while the periphery of the array is grounded to V_{gnd}. (We use V_{fig} = 3.5V and V_{gnd} = 2.0V.) If the contour is complete, the voltage at each interior point rises to V_{fig}, while all outside grid points will settle to V_{gnd}. Thus, the object is rapidly segregated from the background independent of the complexity or the arc length of the contour. If the contour is broken, the saturating resistors (indicated in Figure 1a with simple resistors) will limit the current flowing through these holes in the contour and partially seal off the boundary.

Figure 1b represents a conceptual view of how an object

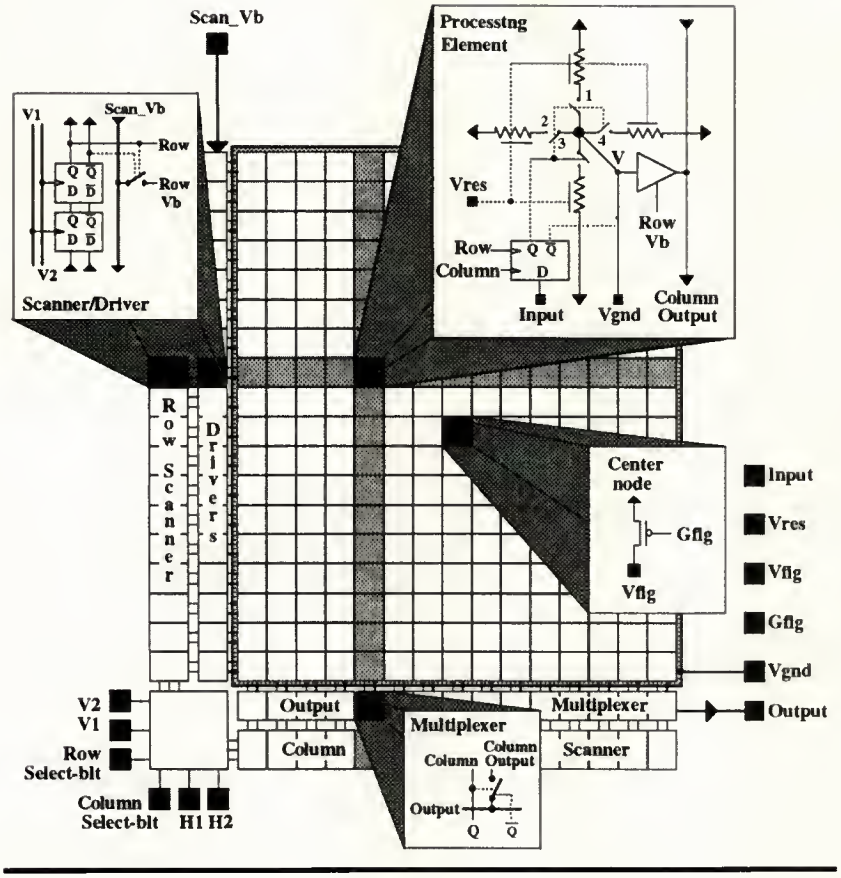


Figure 2. Circuit floor plan.

(figure) is segregated from the background in the 2D view field, in terms of two distinct voltage levels (V_{fig} versus V_{gnd}).

Contours in real images are frequently incomplete and contain broken segments or gaps of one or more pixel widths. As a result, the current flows through these "holes" in the contour, smearing the voltage level between inside and outside. Performing an additional processing step after edge detection, known as edge completion, can, in principle, close the incomplete contours. Currently, there are no efficient digital-machine vision algorithms to find and close incomplete edges. Furthermore, such digital algorithms are expensive to implement in electronic circuits.

We use a different route toward boundary completion by exploiting a property of the saturating resistance circuit, HRes (refer to HRes box again). Its I-V relationship is linear for a small voltage range around the origin; for large voltage gradients across this device, the current saturates at the current I_{sat}. At those locations in the Figure-Ground chip where the contour is broken, the voltage gradient is large, and the saturating resistances limit the current flow, preventing smoothing of the voltage profile.

Implementation

Figure 2 is the floor map of one of our Figure-Ground chips. On this chip the network is formed with HRes as the resistive element. The chip consists of a two-dimensional 48×48 array. We made this version to fit the standard die size (4.6×6.8 mm) provided by MOSIS prototype services. We use the 2.0 µm, p-well, double-metal CMOS process for fabrication of prototypes. No special processes are involved, and we have designed other chips that have larger sizes. We received 12 chips back from MOSIS, eight of which are fully functional.

The Figure-Ground chip is primarily made up of two parts: the resistive network composed with processing elements attached to every node, and a scanning frame that interfaces the network with a computer-controlled data exchange system.

In Figure 2 an array of processing elements is mapped onto a 2D square lattice. Two of the close-up views show the details of processing elements and the current injection at the center node of the network. Row and column scanners appear on the left and at the bottom. Drivers and multiplexers (also shown in close-ups) attaching to scanners serve to access the element array sequentially. Global wires running across the entire network (not shown here) provide biases to processing elements.

Figure-Ground processing elements. The resistive interconnections are actually implemented by HRes pass transistors, which are biased by shared HRes bias circuitry. In addition, four switches (in series with pass transistors) surround the node. These switches are controlled by the input bit, which is stored in the set-reset logic element (described in the Two Key Elementary Circuits box).

If a contour in the image crosses a pixel, the input data Q, which is stored in the set-reset logic element, turns off the switches in the processing element corresponding to the pixel, isolating this node from its neighbors. A fifth switch, controlled by the complementary of the input signal, serves to ground the node (connects it to V_{gnd}, which represents the background) while the node is isolated from its neighbors. If the contour input does not appear at the pixel, the four switches remain closed, and the fifth opens; thus the local horizontal connection of the network is completed at this node.

Two key elementary circuits

Our Figure-Ground chip contains two key elementary circuits: a transconductance amplifier[10] and complementary set-reset logic (CSRL).[18] Figure C1 shows the output current as a function of the differential input at V_1 and V_2. The differential current flowing through T_1 and T_2 has a form $I_1 - I_2 \propto I_b \tanh[(V_1 - V_2)/2]$, where I_b is the current biased by T_b, a single-transistor current source. A transconductance amplifier is frequently used in our systems as followers, that is, to connect its negative input to its output.

Transconductance amplifiers can operate either in current or voltage output mode, depending on the impedance of the load, which varies according to the circuit architecture. For example, a surface interpolation chip uses a transconductance amplifier as current injection elements (see the Resistive Networks box, next page). The input voltage is converted into current injected into the network node. The bias of the amplifier, which controls the current gain of the amplifier, represents the conductance G. Another example is to use it as a voltage follower to buffer analog signals. A voltage follower can be used as a buffer in a sample-and-hold circuitry as is used in the surface interpolation chip (see the Resistive Networks box), or as an analog buffer to gate the signal or to convert the voltage signal into a current signal. In either case, the bias transistor acts as a triggering device. The amplifier operates over nearly the whole voltage range from ground to near V_{dd}. A wide-range version (not shown in the figure) is also frequently used in applications that critically require full-range operation.

In Figure C2 the complementary set-reset logic element functions as a bit storage element. Our chip uses it to store the edge input in the processing elements and as shift register elements in the scanners. The main body of the CSRL is a pair of cross-coupled inverters. Complementary inputs are required. The circuit is simple and efficient. We find it is easy to fit a CSRL element in the processing element even though layout area is usually very limited.

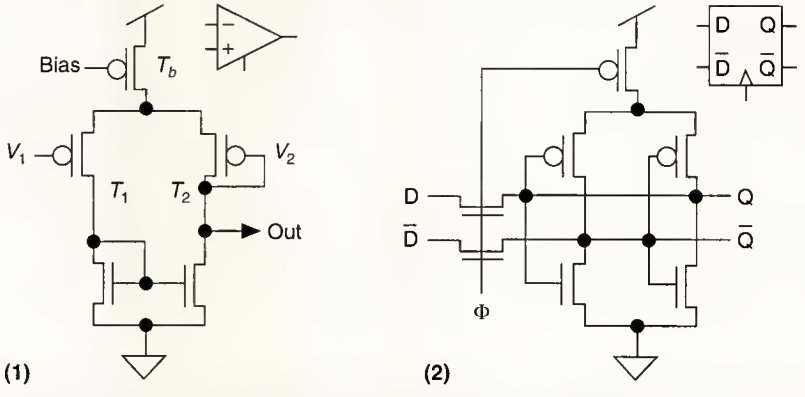


Figure C. Schematic diagrams of transconductance amplifier (1) and the complementary set-reset logic element (2).

To read voltage V at the network node, a column output line runs through the processing element and connects outputs of all the processing elements in the same column. An analog buffer, made up of a transconductance amplifier, is used for the voltage output. When a row selection signal (Row V_b) appears on a row select line that triggers the buffer, node voltage V is duplicated on the column output line. Each processing element is laid out with an area of 100×74 μm and includes 30 CMOS transistors. The same processing element is used at every network node except the one located at the center.

Current injection at the center node of the network. The element at the center node of the network is shown in another close-up view in Figure 2. The key element in the center node is a single-transistor current source. The voltage V_{fig}, representing the figure, is applied onto the center node through the transistor. This fairly large (in terms of channel width) transistor provides sufficient current to counteract the effect of small leakage current throughout the network. Bias G_{fig} controls the conductance G of the current injection to the network (see the Resistive Networks box, next page). This current source configuration allows us to adjust the strength of the injected current. This is useful for chip characterization.

Input configuration. To sequentially access all processing elements on the entire 2D grid, we constructed an on-chip 2D scanning frame. Shift registers are used in both the row scanner and column scanner. A pair of set-reset logic cells (described in the earlier Two Key box) makes up a single scanner stage. (See another of the close-up views in

(continued on p. 53)

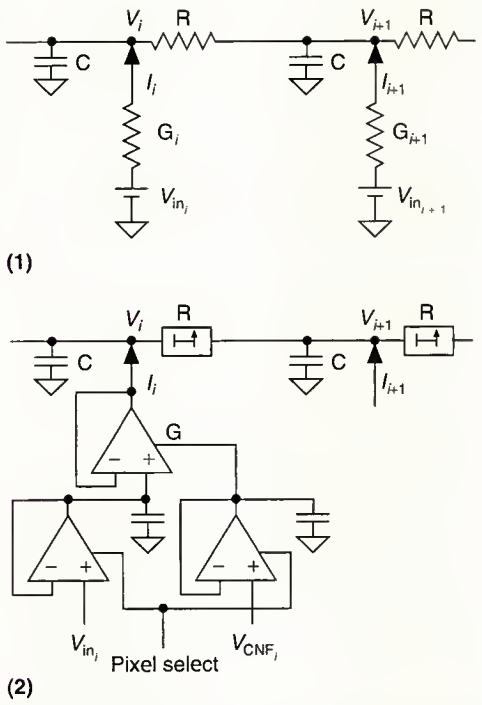
Resistive networks

Figure D1 illustrates a basic resistive network for analog computation of early vision algorithms. Figure D2 shows a typical CMOS analog circuit needed to construct the resistive network for a 2D surface interpolation chip.[19] This circuit helps reconstruct a sparsely sampled and noisy surface. The operation of the chip is based on smoothness and the segmentation assumptions.[8] We use it here to demonstrate the basic properties of our silicon resistive networks.

Two input data arrays, appearing at a node as V_{IN_i} and V_{CNF_i}, are electrically scanned onto the resistive network. At each node of the network, the voltage supplied at G_i corresponding to the "confidence" in the value V_{IN_i} at the ith node, controls the conductance between the intensity input V_{IN_i} and the network node. In case the data at a pixel is invalid or absent, a zero-confidence signal supplied to the pixel shuts off the data path of the input to the network node. V_{CNF_i} sets the conductance G to a certain level according to the level of the confidence the user has in the input at that location. For data corrupted by additive Gaussian noise of variance σ^2, the value of G, that is, the "confidence," is set to $1/(2\sigma^2)$. The value of the horizontal resistance R is electrically adjusted via a common bias wire so that the space constant of the entire network can be changed globally.

Two subthreshold operating circuits operate as key elements of the network.[10] HRes acts as a horizontal resistive element of the network, and a transconductance amplifier acts as a vertical current injection element. The resistive network implements smoothness in areas of the image where the spatial gradient of input intensity is small, and HRes can therefore operate as a linear resistance (see earlier HRes box).

Figure D3 illustrates the empirically measured Green Function response of the circuit, that is, its response to a single pixel being set to a constant voltage (here to 2.4V). This function has the form $V_o e^{-|X|/\chi}$, where χ is the space constant, with $\chi = \sqrt{(RG)}$. These measurements were taken from a 48×48-pixel, 2D resistive network (solid dots). We fit the function $V(i) = 2.1 + 0.3\, e^{-|24-i|}$ through these data points (solid line), where i is the node number. The close match between experiment and theory is evident.



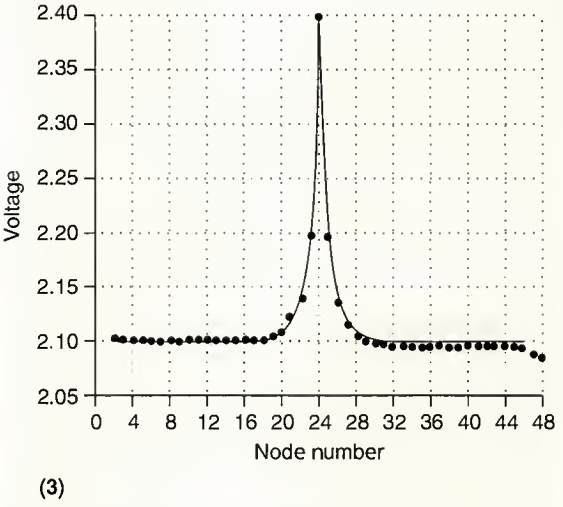


Figure D. One-dimensional resistive network (1), its circuit implementation (2), and the voltage in response to a current injection at node number 24 (3).

Figure 2.) Using a nonoverlap, two-phase clock, a select-bit can be shifted along the shift register (scanner). During the operation, select-bits are shifting in the row and column scanners. They select one row and one column of network nodes during every clock state. At the intersection of the selected row and column lines, only one processing element is selected at a time.

Data exchange. Off the chip, a general-purpose data-acquisition interface card and a software state machine exchange data between the chip under test and a personal computer. The interface produces both vertical and horizontal clock states (V1, V2, H1, H2) and select-bits (row select-bit and column select-bit). One row and one column is selected at each clock state. At each state, the PC interface also writes the image data to the selected processing element and reads the output from it. The entire array of the elements is repeatedly scanned so that the updated responses can be observed continuously on a PC screen. See schematics of the scanner/driver and multiplexer in Figure 2.

The test pattern (a binary pixel map) and the response (an analog data map) can be interactively viewed, analyzed, edited, or stored on a PC. The programmable software supports different chip versions, which have different sizes of the data array, different input/output data configurations, and/or different polarities of clock signals, and so on.

Performance

Figures 3 and 4 (on the next page) illustrate the performance of the Figure-Ground chip. If the contour is unbroken, the voltage inside the figure rises to V_{fig}, segregating it from the surrounding area. If a small gap appears in the contour, it can be partially "sealed off" by the action of the saturating resistance HRes, limiting the current flowing through this gap and thereby inhibiting full voltage equalization from occurring across the break.

As the break in the contour becomes larger, the voltage gradient becomes smaller and smaller and the chip fails to discriminate unambiguously between "inside" and "outside." Yet for a small-enough break along a continuous contour, humans tend to perceive illusory contours, completing the contour even though no real edge exists at the location of the break. It is, however, somewhat arbitrary at what distance two aligned edges are considered to be part of the same or separate contours (Figure 4). If the global threshold is set to 3.0V (in the case of Figure 4b), the contour with one or three pixel breaks would be considered a single figure, while the two larger breaks would not be.

In Figure 3 the responses of the Figure-Ground chip to different input patterns are collected with a fixed-bias set: V_{fig} equals 3.5V, G_{fig} equals 2V, V_{gnd} equals 2V, and the HRes bias V_{res} equals 4.3V. We show the 2D data as pairs of images. The input patterns are located on the left, while the corresponding voltage outputs appear on the right. The black-

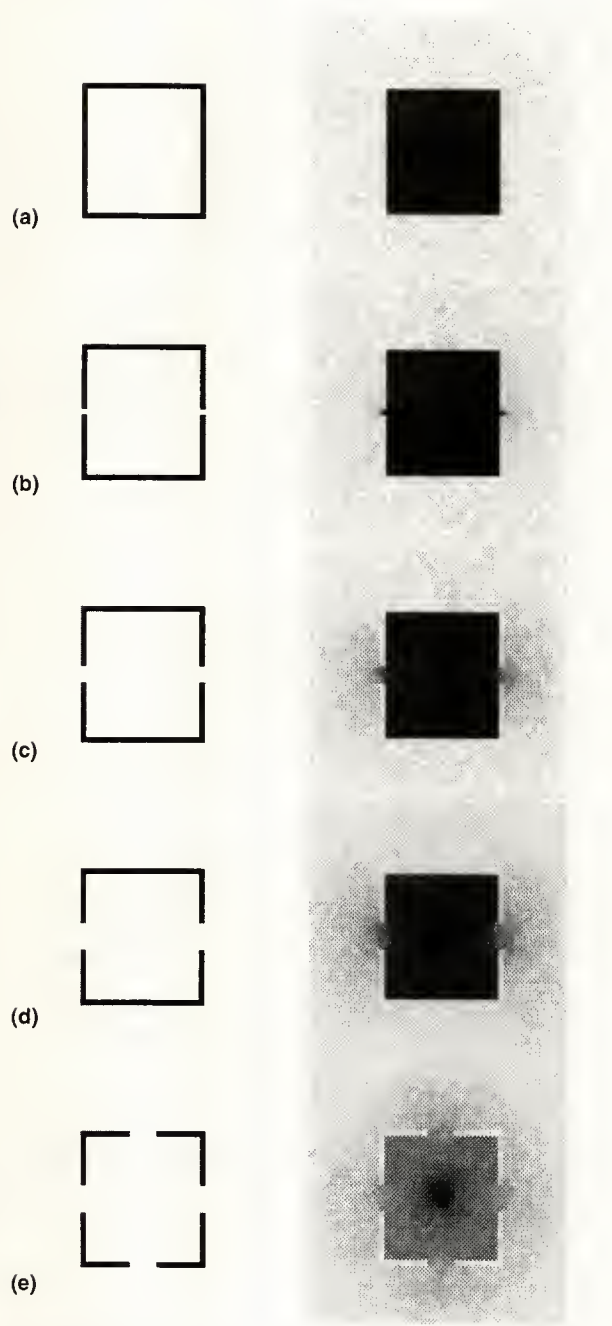


Figure 3. Measured responses to different input patterns: a completely closed box contour (a), the box with two 1-pixel-wide breaks (b), with two 3-pixel-wide breaks (c), with two 5-pixel-wide breaks (d), and two additional 5-pixel-wide breaks (e). The left side of the figure shows input patterns; the corresponding voltage outputs of the chip appear on the right.

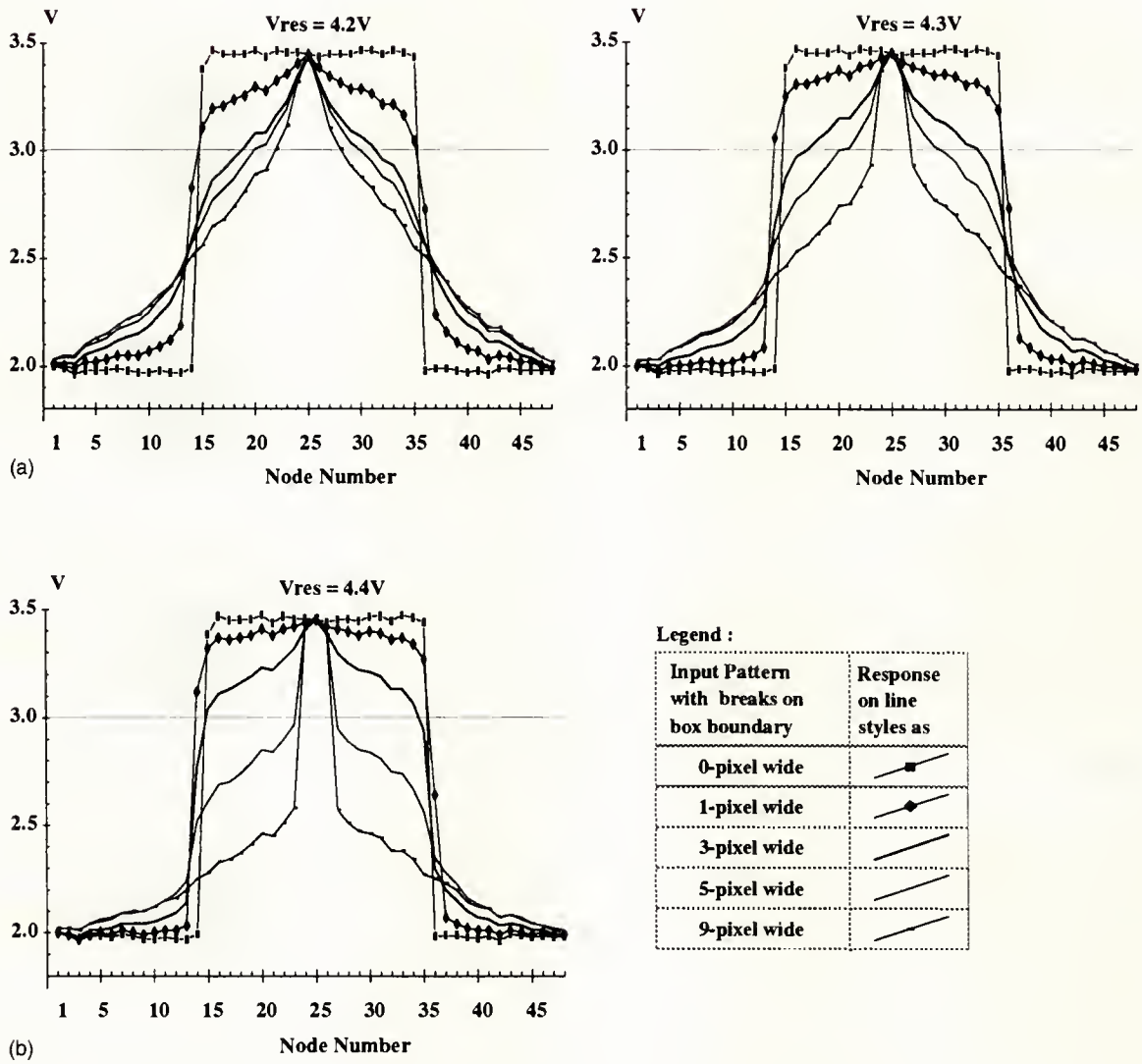


Figure 4. Cross section of voltage along the central row for different resistance values of HRes: low (a), medium (b), and high (c). Taken from Figure 3.

white patterns represent the binary input data encoding object boundaries. Thus, at all locations marked in black, the associated switches shown in Figure 1a are opened. The gray-scale denotes output voltage levels, with the darkest value corresponding to V_{fig} and the brightest to V_{gnd}. The center pixel of the view field is always set to V_{fig}.

In Figure 3a the input consists of a completely enclosed box. The network is therefore broken into two isolated segments, the inside and the outside of the box. Figure 3b shows the object boundary with a break equal to one pixel at the center of the left and right edges. Due to the large voltage difference across these two leaks, the saturated HRes horizontal resistances saturate, thereby helping to seal off these breaks by exploiting the saturating properties of HRes.

In Figure 3c the width of the breaks in the contour increases to three pixels each. Yet HRes still acts to effectively seal the two holes, and the figure is segregated from the surrounding areas. In Figure 3d the width of the breaks increases to five pixels each. Due to the much smaller voltage gradient across this wider gap in the contour, the voltage

spreads outside the figure. In Figure 3e a total of four breaks, each five pixels wide, prevents the figure from being segregated at all. The system can't decide whether a single object with wide breaks along its side or four separate objects are present.

Note that at every node where a boundary input signal (in black) appears and the switches are opened, the output voltage at that node is tied to V_{gnd}. This can best be seen in the white outline in Figure 3e. For small-enough breaks, our circuit has an excellent boundary-completion capability. This is important for machine vision, since real images rarely have complete boundaries.

Figure 4 plots the voltage profile across the chip in response to the same set of complete and broken square contours as in Figure 3 for different values of HRes. Of the five curves in each of the three plots, the first four correspond to Figure 3a-d. The contour with four breaks in Figure 3e has been replaced with two breaks of nine pixel widths each. The boundary is always located at pixels 14 and 36 and center-row symmetrical. The 1D voltage profiles shown are taken from the center row of the data array, where the breaks occur.

HRes is biased to 4.2V, 4.3V, and 4.4V in the three plots, corresponding to a low HRes value in Figure 4a, medium in Figure 4b, and high in Figure 4c. The center pixel (25) is set to V_{fig} (here 3.5V), and the nodes along the edges of the network (pixels 1 and 48) are always set to V_{gnd} (here 2V). The voltage response of the circuit to the completely closed boundary box (top curve) is very close to V_{fig}, and the voltage profile is flat in the remainder of the network. As the width of the two breaks in the contour increases from one to nine pixels, the voltage profile across the break, at first very steep, becomes less steep and flattens out eventually. Furthermore, as the value of the horizontal resistance increases (going from Figure 4a to 4c), the voltage gradient across the break becomes steeper for smaller breaks, improving the contour-completion capabilities of the circuit. (Compare the top three curves in all three plots.) If all pixels with an output voltage above 3.0V are considered to belong to the figure, the unbroken figure and the figure with 1-pixel-wide breaks (as well as the box with 3-pixel-wide breaks in plot Figure 4c) would have been segregated. In the case of the 5- and 9-pixel-wide breaks, the voltage roughly decays exponentially with distance.

Figure 5 demonstrates the Figure-Ground chip's response to a real image of a moving hand.[19] We preprocessed the raw video image to yield a set of noisy edges outlining the hand. We then scanned these edges onto the Figure-Ground chip (Figure 5a). The output data (Figure 5b) is shown in 3D plots in which the vertical dimension represents the node voltages. Furthermore, we shaded all pixels in Figure 5c that have an associated node voltage above 2.4V. (The central point at which the voltage equals V_{fig} is indicated in black.) Such a simple decision rule successfully labels all pixels as-

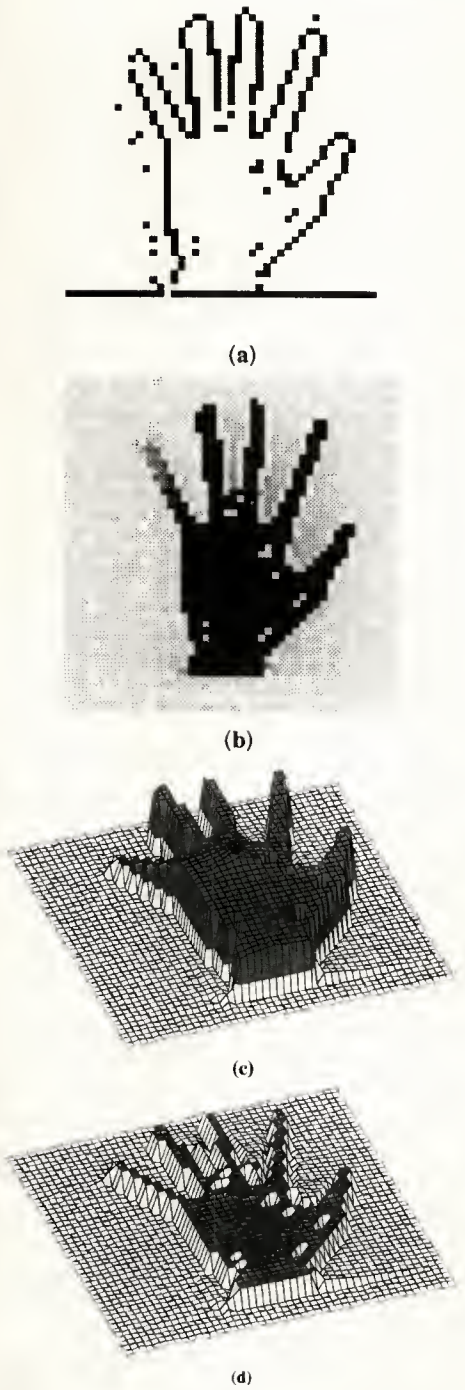


Figure 5. Experiments on a real video image of a hand: binary edge map input (a), output of the Figure-Ground chip (b-d). The intensity representation in (b) is converted into a 3D plot in (c). All points above 2.4V are indicated in gray. This figure is thresholded at 2.4V (d).

sociated with the entire hand, as shown in Figure 5d, despite the gaps in the outline of the hand. Note that the voltage decays rapidly along the little finger (Figure 5c) because the finger tip has an incomplete contour (Figure 5a). However due to the saturating HRes operation, a considerable voltage drop, sufficient to segregate the little finger from the background, still occurs. The HRes bias V_{res} equals 4.4V; all other values are as shown in Figure 4.

The white spots inside the hand (corresponding to the "holes" in the 3D plots) as well as the white contour surrounding the hand occur when the chip assigns edges in the image to nodes in the resistive network and pulls them to V_{gnd}. A newly designed circuity will avoid this problem by mapping the contour onto the resistive connection between nodes.

OVERALL, THE FIGURE-GROUND CHIP behaves very satisfactorily. In particular, it performs figure-ground segregation in the presence of incomplete and broken boundaries, an ever-present feature of real images. Its boundary-completion capability is due to the saturating nature of HRes and does not depend on any complicated, nonlocal machine-vision type of algorithm. Our system can therefore replace a considerably more complex set of digital algorithms with a single dedicated analog circuit. Applications of this circuit include situations in which the rapid identification of a target from a cluttered background is essential. The object, once segregated from the surrounding areas, can be further processed for identification of other tasks.

The two major limitations of the current Figure-Ground chip are its limited capability for recognizing figures with large gaps in the contour and the constraint that the figure always has to be centered. We are now designing circuits replacing the saturating resistances HRes with resistive fuses[8,9] for even better "contour-sealing" performance. (In a resistive fuse the current goes to zero rather than to a constant value when a large voltage difference is applied across this circuit.) Furthermore, multiple, selectable, current injection nodes will enable us to select any figure in the scene for labeling. This is somewhat analogous to a spotlight of attention.

An additional modification will increase the spatial precision of the contour representation by allowing edges in the image to map onto the resistances connecting adjacent nodes, rather than to the entire node as in the current version of this chip. We have implemented these changes, and a chip is under fabrication.

What is the outlook for such analog CMOS ASICs for early and intermediate vision? Given the low cost and small size associated with these chips, they clearly fill a niche in a host of military, industrial, and household applications, in particular, for surveillance and tracking applications. Apparently, real-time, small, power-lean, and robust analog computers are making a limited comeback in the form of highly dedicated, smart vision chips.

Acknowledgments

The National Science Foundation, the Office of Naval Research, and Rockwell International Science Center supported the circuit design and fabrication in the laboratory of Christof Koch. Rockwell International Science Center also supported the research and development work conducted by Tanner Research, Inc. Tanner Research developed and provided the mask layout design and verification tools. We thank John Harris and Shih-Chii Liu for the technical discussions.

References

1. "Visual Information Processing: From Neurons to Chips," *Proc. SPIE*, B. Mathur and C. Koch, eds., Vol. 1473, 1991.
2. C. Koch, "Seeing Chips: Analog VLSI Circuits for Computer Vision," *Neural Computation*, Vol. 1, 1989, pp. 184-200.
3. Special Issue, *Int'l J. Computer Vision*, J. Clark, ed., Vol. 7, No. 3, 1992.
4. B.K.P. Horn, "Determining Lightness from an Image," *Computer Graphics and Image Processing*, Vol. 3, 1974, pp. 277-299.
5. T. Poggio and C. Koch, "Ill-Posed Problems in Early Vision: From Computational Theory to Analogue Networks," *Proc. R. Soc. Lond. B*, Vol. 226, 1985, pp. 303-323.
6. M. Bertero, T. Poggio, and V. Torre, "Ill-Posed Problems in Early Vision," *Proc. IEEE*, Vol. 76, 1988, pp. 869-889.
7. T. Poggio, V. Torre, and C. Koch, "Computational Vision and Regularization Theory, *Nature*, Vol. 317, 1985, pp. 314-319.
8. J.G. Harris, C. Koch, and J. Luo, "A 2D Analog VLSI Circuit for Detecting Discontinuities in Early Vision," *Science*, Vol. 248, 1990, pp. 1209-1211.
9. J.G. Harris, C. Koch, J. Luo, and J. Wyatt, "Resistive Fuses: Analog Hardware for Detecting Discontinuities in Early Vision," *Analog VLSI Implementations of Neural Systems*, C. Mead and M. Ismal, eds., Kluwer, Normwell, Mass., 1989, pp. 1-29.
10. C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley, Reading, Mass., 1989.
11. C.A. Mead, "A Sensitive Electronic Photoreceptor," *Proc. Chapel Hill Conf. Very Large Scale Integration*, H. Fuchs, ed., Computer Science Press, Rockville, Md., 1985, pp. 463-471.
12. M.A. Sivilotti, M.A. Mahowald, and C.A. Mead, "Real-Time Visual Computation Using Analog CMOS Processing Arrays," *Proc. Stanford*

Conf. VLSI, MIT Press, Cambridge, Mass., 1987, pp. 295-312.
13. S. DeWeerth and C. Mead, "Analog VLSI Circuits for Stimulus Localization and Centroid Computation," *Int'l J. Computer Vision*, Vol. 8, 1992, pp. 181–202.
14. D.L. Standley, "An Object Position and Orientation IC with Embedded Imager," *IEEE J. Solid-State Circuits*, Vol. 26, 1991, pp. 1853-1859.
15. S.C. Liu and J. Harris, "Dynamic Wires: An Analog VLSI Model for Object-Based Processing," *Int'l J. Computer Vision*, Vol. 8, 1992, pp. 231–239.
16. S. Ullman, "Visual Routines," *Cognition*, Vol. 18, 1984, pp. 97-159.
17. W. Bair and C. Koch, "An Analog VLSI Chip for Finding Edges from Zero-Crossings," *Advances in Neural Information Processing Systems*, Vol. 3, D.S. Touretzky and R. Lippman, eds., Morgan Kaufmann, San Mateo, Calif., 1991, pp. 399-405.
18. C. Mead and J. Wawrzynek, "A New Discipline for CMOS Design: An Architecture for Sound Synthesis," *Proc. Chapel Hill Conf. Very Large Scale Integration*, H. Fuchs, ed., Computer Science Press, 1985, pp. 87–104.
19. J. Hutchinson, C. Koch, J. Luo, and C. Mead, "Computing Motion Using Analog and Binary Resistive Networks," *Computer*, Vol. 3, 1988, pp. 52-63.

Jin Luo is a staff scientist at Tanner Research, Inc., Pasadena, California. He designs analog ICs for neural networks, vision, and a variety of ASIC development projects. Previously, he worked as a staff research assistant in the Computation and Neural Systems Program at the California Institute of Technology in Pasadena.

Luo received his MS degree in computer science from the California Institute of Technology, Pasadena.

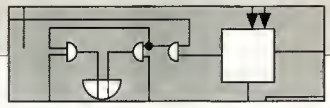
Christof Koch is associate professor of computation and neural systems at the California Institute of Technology, where he studies how single brain cells process information and the neural basis of motion perception, visual attention, and awareness. He also designs analog VLSI vision chips for intelligent systems.

Koch was awarded his PhD in biophysics by the University of Tubingen in Germany.

Bimal P. Mathur, manager of the Analog Image Processing Department at Rockwell International Science Center, is currently investigating parallel architectures for early vision procession. In the past, his research interests included analog signal processing devices (CCDs and SAWs).

Mathur holds a BE degree from Birla Institute of Technology and Science, India, and an MS and ScD from the Massachusetts Institute of Technology, Cambridge.

Direct questions concerning this article to Jin Luo, Tanner Research Inc., 180 North Vinedo Avenue, Pasadena, CA 91107; or e-mail at jin@tanner.com.



The Associative Processor System CAPRA : Architecture and Applications

Associative processor systems are of growing interest in certain application fields. The innovative features of the novel architecture that we propose for such a system include intelligent memory cells (they directly include processing logic), a maskable memory decoder supporting multiaccess operations on the array, and integration of optical sensor elements. We describe the basic features of our content-addressable processor/register array (CAPRA) and discuss its potential for applications in database support, basic numerical tasks, and image processing.

Karl E. Grosspietsch

German National Research Center for Computer Science

Ralf Reetz

University of Karlsruhe

Advanced hardware integration techniques like very large scale integration (VLSI) or wafer scale integration (WSI) imply the potential to efficiently implement new architectures formerly unrealized because of technological restrictions, such as pin limitations or too-small bit capacities. In this context, approaches that eliminate the bottleneck between processor and data appear especially interesting.

In conventional von Neumann machines, data must be fetched from memory and transferred to the processor every time it is manipulated. The result is then stored into memory. So, for many of these operations, data transfers account for most of the execution time.

One solution to the processor/memory bottleneck is to integrate more logic directly into the memory structure—that is, to make the memory more intelligent. Such intelligent-memory architecture especially applies to nonnumerical data processing fields like database management, logic programming, pattern recognition, image processing, and CAD graphics.

Because they are a step toward smarter memories, systems for associative (meant here as a synonym for content-addressable) data processing can again become important.[1] For the first time, hardware integration promises the implementation of such systems with a reasonable size and cost/bit ratio.

The architectural approach

Several interesting approaches for content-addressable processor systems have been reported in the last few years. The June issue contains a comprehensive survey.[1] We base our approach mainly on the ideas of Lea,[2] extending that solution to achieve the following objectives:

- increase the flexibility of logic elements,
- combine processor cell arrays with ordinary content-addressable memory (CAM) and RAM parts, and
- modify the resulting architecture for testability and fault tolerance features.

The latter implies not only structural redundancy (by spare components) but also functional redundancy, in the sense that a more complex component's function can be stepwise degraded to less comfortable functionality.[3]

Basic principles and requirements. We achieve our goals by including a CAM segment in an existing RAM structure. As shown in Figure 1, a RAM, a CAM, and a content-addressable processor/register array (CAPRA) together form a kind of storage hierarchy where the main part consists

0272-1732/92/1200-0058$03.00 © 1992 IEEE

of an ordinary RAM. The "smarter" components are included as additional memory segments within one uniform physical memory space; we can thereby arbitrarily tailor their individual storage capacities to the application's specific needs.

Compatibility between the steps of a hierarchy is achieved in the sense that smarter parts also provide the full functionality of simpler parts. So, the CAM parts are also made operable as RAMs, and some functional extension of the CAM architecture implements the CAPRA.

For the introduction of more flexible logic, the following architectural properties appear promising:

- *Extension of the conventional simple-bit mechanisms of present CAM architectures—equivalence between data, potentially modified by some kind of masking—to solutions that also allow more sophisticated bit evaluation for pattern matching.* (For example, we can use a threshold number of identical bits in the search pattern and compared data to decide about hit, or use other similarity metrics between the search pattern and data in storage.)
- *Extension, with relatively low hardware effort, of the usual comparison logic of a CAM cell to provide an entire set of 1-bit Boolean operations.*
- *Modification of arithmetic logic units from sequential 1-bit adder elements per word cell to at least 4-bit adder elements.* Each of these ALUs has a data path to neighbor ALUs in the two nearest word cells. So, in addition to parallel processing of data, the architecture provides parallel exchange of data between word cells of the CAPRA segment. Apart from requirements to restrict additional area size if possible, we chose a four-bit ALU length to support the processing of pixels with 16 gray levels in image processing.
- *Implementation of features for multiaccess write operations[1,4] and parallel evaluation of test outcomes in word cells using extended logic[3] to support test of the architecture.*

We can easily integrate the proposed architecture into a conventional system because, unlike other unorthodox architectural approaches, our features comply with the von Neumann machine's usual control-flow programming paradigm. We therefore planned our architecture to work as a coprocessor of a conventional main processor; the coprocessor's instructions are modularly added to those of the processor.

The resulting hardware architecture. The following architectural features fulfill our requirements:

- The CAM has an additional RAM access mode so that it can, for example, be loaded or read like ordinary RAM cells.

Frequently used symbolic abbreviations

f	Gray-level intensity
i, j, l, x, y	Index variables
k	Number of pixels stored in a word cell
L_A, L_B	Bit length of records of relations A,B
m	Dimension of vector
M	Dimension of neighborhood matrix
n	Memory word length
N	Dimension of pixel array
N_A, N_B	Cardinalities of relations A,B
P_A, P_B	Smallest power of $2 \geq N_A, N_B$
r	Pixel resolution
w	Word capacity of the CAPRA segment
z	Number of pixel rows stored in CAPRA

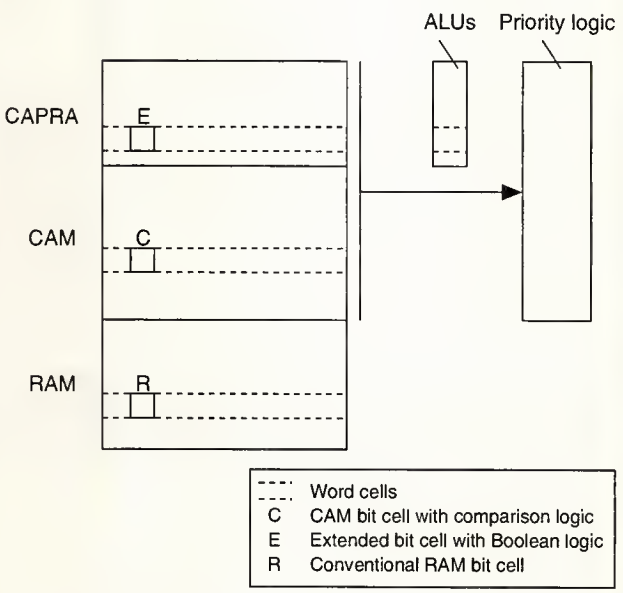


Figure 1. The architecture combines a RAM structure, a CAM segment, and a CAPRA.

- In the CAPRA architecture, illustrated in Figure 2, next page, RAM bit cells again serve as base cells. Moreover, a simple logic block is associated to every bit cell, which enables Boolean 1-bit operations between two 1-bit operands. This allows bit-parallel and word-parallel execution of a Boolean operation on all words of the CAPRA segment.
- In addition, in the CAPRA word cells the classical CAM

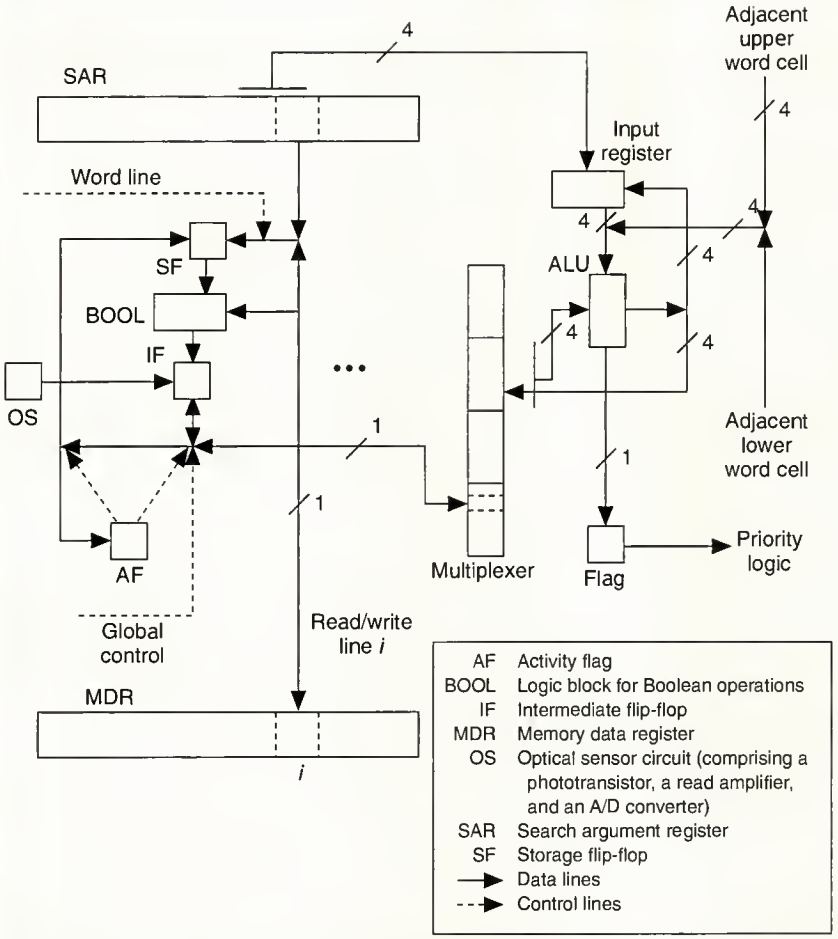


Figure 2. Schematic of the CAPRA architecture showing one extended bit cell *i* (*i*=0, ..., *n*–1) of a word cell, together with the ALU and the optical-sensor element associated to that word cell.

equivalence check operation is possible, realized by three additional transistor functions (refer to the survey article[1]).

- To every word cell in the CAPRA, we assign a simple 4-bit adder/shifter unit; so, for example, with a classical word length of n=32 bits, an arithmetic operation can be performed on all words of the CAPRA segment in parallel in about eight cycles.
- An additional flag bit is associated to each bit cell of the CAPRA, which enables us to flexibly define arbitrary "activity patterns" for the array cells. We therefore can process data not only on all processor elements but also on a previously defined, arbitrary subpattern of processing elements.
- An additional mask register provides a simple extension of the memory decoder[4] for RAM access. By setting bits of this register to 1, an arbitrary part of the address bits can be declared don't-care bits. Thus we can implement concurrent access to word cells that have common address bit subpatterns (the survey article contains a detailed description).

Figure 3 shows CAPRA's operation. Data can be written into a word cell—controlled by the corresponding word line emanating from the memory decoder—via the memory data register and read/write lines. We can combine the contents of every bit of a word cell, contained in the storage flip-flop (SF), with the contents of the read/write line by a Boolean operation in the functional block BOOL. The result of this operation is latched in the intermediate flip-flop (IF). From there it can be propagated further (indicated by control line TRANSFER), either to the adjacent ALU (line LOCAL/GLOBAL = 1) or to be memorized in the SF (LOCAL/GLOBAL = 0). These transfers take place either unconditionally (control line UNCOND=1) or conditionally (line UNCOND=0), depending on the status of the cell memorized in the activity flag (AF). (Signal line COND equals 1 if AF stores a 1.) As a third sink for the bit transfer from the IF, setting of the AF is possible (control line SET FLAG). This is performed either unconditionally (control line UNCOND = 1) or dependent on the present status of the AF (control line COND′ = 1). In the latter case, the AF can be set only if it has not yet been set—that is, if the AF is storing a 0. (COND′ then equals 1.)

The intermediate flip-flop can receive a data bit not only from the functional block BOOL, but, alternatively, from the adjacent ALU (control line REC = 1).

Sensor integration. For image processing applications, we plan to integrate optical sensors with our intelligent bit cell array on one piece of silicon. One sensor element is associated to each word cell of the CAPRA segment. The sensor element comprises a phototransistor, a read amplifier, and a programmable analog-to-digital converter (refer back to Figure 2).

The phototransistor was implemented as a PMOS transistor with floating bulk,[5] an efficient way to integrate it into a CMOS process without process modifications.[6] The

phototransistor's sensitivity depends on its operating point and on the incident light's wavelength.

The converter transforms the incoming analog signal into a digital bit pattern in a number of iterations. The accuracy of the conversion—the bit length of the digitized signal pattern—depends on the number of conversion cycles, so this resolution is easily programmable. We have selected a resolution of r=4 bits for our planned applications. The resulting bits are stored in r consecutive bit cells of the corresponding memory word, starting from a previously determined bit position.

Nearly all approaches for such sensors have been based on analog solutions. But most of these analog converters are not compatible with standard digital CMOS processes—they depend on special complicated fabrication steps that cannot be integrated into the production of standard CMOS structures. As an alternative, we use a sensor element that is fully realized in CMOS technology.[6] Thus we can integrate all the components on a single chip or wafer by one standard CMOS fabrication process.

To realize an A/D converter of programmable resolution, we used the so-called cyclic conversion technique.[7] The converter performs an r-bit conversion in $3r$ clock cycles. This cyclic converter's operation is based on recirculating the input voltage, thus precisely doubling the voltage.[8] For example, a conversion time of 20 µs is necessary for the chosen resolution of 4 bits.

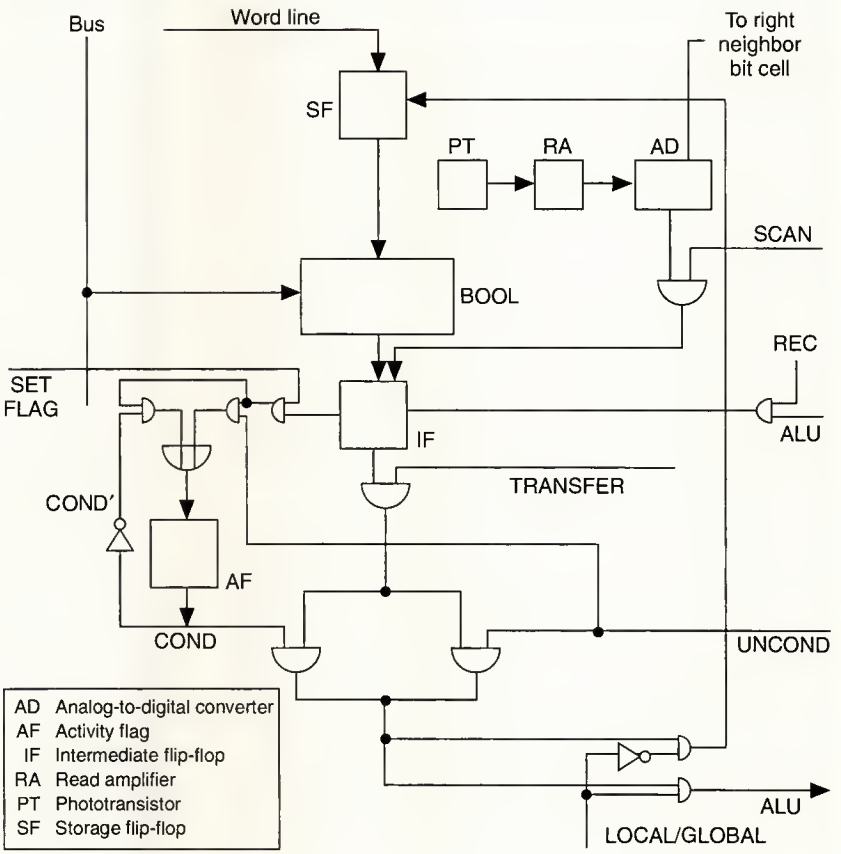


Figure 3. Gate structure of the extended bit cell.

CAPRA's basic instruction set. We have defined the following set of operations for the described memory structure:

- WRITE, ADR; / normal RAM write access (executable in all system parts)
- READ, ADR; / normal RAM read access
- MWRITE, ADR, MASK; / masked RAM write access: multiple access to a set of word cells in memory that have some address subpattern in common
- ASSOCOMP; / word-parallel and bit-parallel comparison of the contents of word cells with a predefined search pattern in the search argument register (executable in the CAPRA and in the CAM part)
- BOOLOP; / Boolean operation combining the bits of all memory words with an external operand's bits. (This operation name is a placeholder for the 16 different Boolean operations of two 1-bit operands.)
- STORE, COND (UNCOND); / stores in the bit cells of the CAPRA part the result of the Boolean operation either unconditionally (for all bit cells in the CAPRA) or conditionally (depending on the local activity flags of each cell)
- SET AF, COND (UNCOND); / transfers the contents of the intermediate flip-flop into the AF either unconditionally or conditionally (only for those bit cells where AF=FALSE)
- SCAN(j); / transfers digitized sensor input with a 4-bit resolution from the sensor elements to the IFs of bit slice j, j+1, j+2, j+3.

We can group CAPRA's ALU operations into unconditional and conditional operations. Unconditional operations are executed in all ALUs. Second-class operations correspond exactly in their structure to those of the first class, except they execute in a local ALU only if it has a flag set to TRUE.

Unconditional operations have the structure ALU OP, BUFFER(j), REGISTER, DESTINATION;. The first operand BUFFER(j) is always a 4-bit segment of all memory words; the index j gives its position. As the second operand REGISTER, we may use the registers REGA, REGB, REGC, or SAR. REGA is the 4-bit input register belonging to each ALU in the CAPRA segment. REGB and REGC just represent the REGA register of the upper (lower) neighbor ALU. As an alternative, we can use the least significant 4 bits of the SAR to provide one global 4-bit operand to all the ALUs of the CAPRA. So, by selecting the second operand, we can combine an operand held in a memory word with local data residing in either the corresponding ALU or one of its neighbors (thus enabling communication between neighbors). Or we can combine the operand with a global operand provided from outside. The sink DESTINATION of the operation is either again BUFFER(j) or the register REGA.

The ALU operations are either unconditionally executed on all CAPRA word cells or conditionally controlled by local ALU flags. The setting of these flags depends either on the outcome of certain ALU operations (as is usual in conventional ALUs) or on an explicit instruction from outside. For the latter purpose we have the operation SET, ADR, MASK, which provides—analogously to the masked write operation MWRITE—access to one or several ALUs in one cycle.

This short list omits instructions that deal with handling priority operations and with data transfers between the memory and the main processor.

State of the system implementation. The entire architecture has been specified with the VHDL hardware description language. Not only does this description cover the register transfer level, but we also defined our own abstract data types to exactly model the behavior of our circuits' basic transistor functions at switch level. This level models transistor functions as digital switches. It coarsens the more detailed physical characteristics of the transistor such as delay times, switching speed, and analysis of transient behavior. On the other hand, it provides more refined information about the transistor than the usual gate-level models used to study the steady-state behavior of circuits.[9]

We based our circuit model on a six-valued logic that comprises two different low-impedance states, three high-impedance states, and one undefined/unknown state. Apart from contributing somewhat to the emerging VHDL design technique, the detailed switch level model of our architecture is especially useful for fault simulation and derivation of test patterns.

Together with the simulation environment of VHDL, our description also provides an exact runtime simulator of the specified architecture. In addition to measurements made at the switch level, we aggregated the fine-grain routines of this simulation to higher units at the register transfer level. This allows us to measure performance in units of machine cycles with considerably reduced computing time.

Correspondingly, for the machine language introduced earlier, we wrote an assembler to enable the development of symbolic programs for the described architecture. The assembler transforms symbolic instructions into binary machine words that the simulator interprets. In addition, we implemented a simulator environment that can combine CAPRA machine language procedures with high-level language main programs (to be executed on a main processor) written in Modula. Based on the translator and its environment, a number of application examples currently are being studied and demonstration software implemented.

At the level of the basic hardware circuits, we exhaustively simulated central components (the intelligent bit cell, the ALU part) using the transistor simulator SPICE. Corresponding layouts were generated and partially transformed into silicon.[8]

In the future, we plan to integrate the different developed cell layouts on a common chip.

Database applications

Associative processing is especially useful for applications where data is structured in sets or arrays. Because database applications involve set-like data organization, this field always has been one of the principal applications of associative processing.[10] We illustrate the merits of our CAPRA approach by considering some classical operations in relational databases: selection, intersection, product, semi-join, and join. As a comparison, we refer to an investigation by Fernstrom, Kruzela, and Svensson at the University of Lund, Sweden.[11]

We consider the basic elements of the database, the data records—that is, ordered tuples of data items. Each subfield of a record stores one item. Sets of such records (often called relations) are represented by tables of such records. The relational operations we mentioned work on either one or two tables as input operands, producing a third table as the result. Let us call the source relations A and B. Without loss of generality, we assume that the number of data records in A is the same as or more than those in B. For the cardinalities N_A and N_B of A and B, we thus have the condition $N_A \geq N_B$. Correspondingly, L_A and L_B denote the bit lengths of the records of relations A and B.

In our CAPRA approach, a record is stored in a number of consecutive words in memory (that is, if the record contains more than the n bits fitting into one memory word). Correspondingly, search patterns ranging over the entire bit length of the record must be split up into a number of search words (each of n bits), which subsequently are used for search operations. So, associative checking of the records of relations A or B can be performed in $\lceil L_A/n \rceil$ or $\lceil L_B/n \rceil$ cycles, respectively (with $\lceil x \rceil$ denoting the smallest integer $\geq x$). It is not necessary to store a table—that is, a set of records—in a segment of consecutive data words. Instead, in an associative system, it is possible to characterize the members of the

set by some common properties, namely, values of some items or a common mark bit.

The selection operation. The simplest relational database operation, the selection, selects a subset of the given set A of tuples. This subset comprises records that obey certain search criteria for one or several tuple items. In CAPRA this operation can be carried out by a number of simple associative checks that consecutively compare the values of some tuple fields of the records with search patterns in the SAR; hits are memorized by setting a new mark bit in the data record. Thus, the entire select operation can be carried out by at most $\lceil L_A/n \rceil$ associative search operations; then in one cycle, a mark bit is written into all records found. If no ordering of items with regard to one fixed property is possible, so that hashing techniques cannot be applied, the same search procedure on a von Neumann machine takes the order $O(N_A \cdot \lceil L_A/n \rceil)$.

The intersection operation. This operation has two source relations as inputs and, as a result, produces a common subset of both relations. The CAPRA system carries out this operation by sequentially reading the contents of the smaller relation B. The records of relation A are compared with one record of relation B by $\lceil L_A/n \rceil$ associative search operations. So, processing the entire intersection operation on CAPRA takes the order $O(\lceil L_A/n \rceil \cdot N_B)$, compared with $O(N_A \cdot N_B \cdot \lceil L_A/n \rceil)$ on a von Neumann machine. Analogous time complexities also turn out for the union operation.

The product operation. The Cartesian product operation applied to relations A and B generates all pairs of records. One record of the pair belongs to relation A, the other to relation B. This operation has a time complexity of $O(N_A \cdot N_B)$ on a sequential von Neumann architecture as well as in the bit-sequential, word-parallel LUCAS (Lund University Content-Addressable System) approach of Fernstrom, Kruzela, and Svensson. The CAPRA approach can considerably reduce this time complexity, at the expense of increasing the set cardinalities to the smallest powers of 2 that are $\geq N_A$ ($\geq N_B$); let us call them P_A and P_B. Then, the concatenated pairs of data records can be produced very efficiently in time complexity $O(\lceil L_B/n \rceil \cdot P_A + \lceil L_A/n \rceil \cdot P_B)$. Figure 4 shows that, in one cycle, a masked write access produces P_B copies of a data word of the first relation. These masked write operations are performed for all members of the first relation.

Then, with a changed mask, P_A copies of all the second relation's members are produced in $P_A \cdot \lceil L_B/n \rceil$ masked write operations. Thus, compared to the other mentioned computer architectures, time complexity is reduced by a factor of about N_B at the expense of a memory space capacity of the order $O(P_A \cdot P_B)$. But this seems justified because memory costs are falling drastically.

The semi-join operation. As a characteristic example of this operation, the study by Fernstrom, Kruzela, and Svensson considers two relations, one with attributes g and h, the other with attribute h. The result is a subset of the first relation,

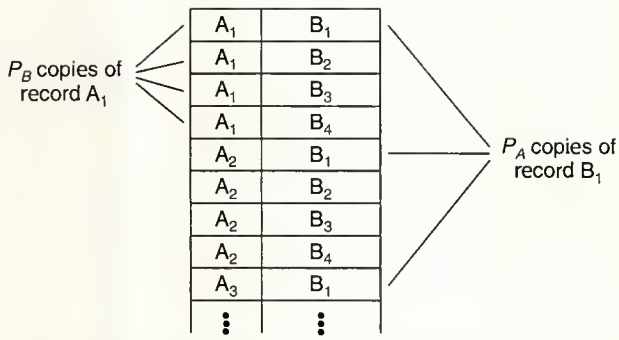


Figure 4. Generation of the two Cartesian product relations A and B using masked write operations. (A_1, A_2, ..., are records of relation A; B_1, B_2, ..., are records of relation B.)

consisting of tuples where the value of attribute h is the same as some value of attribute h in the second relation. This can be produced by sequentially taking members of relation B and associatively checking them against relation A. This takes a time complexity of $O(\lceil L_A/n \rceil \cdot N_B)$.

The join operation. This operation is similar to the semi-join, but matching tuples are concatenated, thereby removing the joining attribute. With CAPRA, we carry out this operation by first sequentially scanning the N_B tuples of relation B. Now relation B is marked, showing where a match occurs in the tuples of relation A. Subsequently, the Cartesian product of the tuples of relations A and B is formed by $\lceil L_B/n \rceil \cdot P_A + \lceil L_A/n \rceil \cdot P_B$ masked write operations. Tuples not marked as matching are logically removed by setting an invalid bit. In the remaining tuples, we remove the attribute to be erased by simply setting another invalid bit in the subfield containing that attribute.

In an analogous way, we can use multiaccess operations to speed up the join operation. Because several (slightly different) versions of this operation have been proposed in the literature,[10] we shall not discuss these modifications here.

In general, CAPRA's parallel search features reduce the time complexity of relational database operations by the order of the cardinality of the larger of the two input relations—that is, by $O(N_A)$ in our example. The CAPRA approach has further advantages if data need not only be found but also updated in some regular way (for example, incrementing or decrementing a subfield in all records found).

Basic numerical operations

We did not specifically intend for our architecture to efficiently support operations of scientific and numerical computing such as matrix-oriented operations (matrix-vector multiplication, matrix-matrix multiplication, fast Fourier transform). But it turned out that the architecture does have some

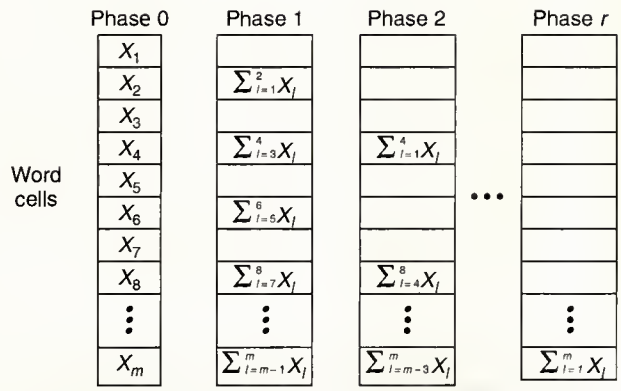


Figure 5. Parallel summing up of the components of an m-dimensional vector.

interesting features in this area.

As a simple example, we use the summing up of the components of a vector of dimension m. A single processor system must sequentially carry out this task in m–1 steps. Our architecture yields a much better result because of the parallelism of computing elements, even though only a simple shift-register-like interconnection joins them.

Consider the vector components to be stored in consecutive word cells, as shown in Figure 5. The first summing-up phase is carried out by transferring each odd-indexed component to its next lower word cell and adding its value to that of the neighbor's cell. So, half the computations necessary for summing up can be executed in parallel, independently of the vector's dimension (provided that this dimension is less than the number of word cells in the CAPRA segment). In the next phase, all the odd-indexed partial results are moved down two word cells to meet the next corresponding partial result. This can be performed in two machine cycles. So, again the partial results have decreased by half in number, while their distances within the CAPRA segment have doubled.

If this strategy is always used, the need to move partial sums over increasing distances sometimes will win over the advantage of being able to sum up partial results in parallel. So, after the described strategy is used for an optimal number of phases, the few remaining partial results (now situated within the CAPRA segment very far away from each other) are then simply read and added sequentially. Reetz shows in detail that if one switches strategy at the optimal point, the entire summing up can be carried out in $O(\sqrt{m})$ steps.[12] For larger values of m, this result is significantly better than the $O(m)$ steps necessary in the case of a single processor. The optimal value for summing up m numbers is of order $O(\log m)$; however, a binary adder tree is necessary for summing up the different values. So, whereas this optimal solution needs complex extra hardware, our architecture just as a by-product performs close to the optimal value for many values of m used in practice.

Reetz[12] shows that, based on this effect, the multiplication of an $m \times m$ matrix with an m-dimensional vector can be performed in order $O(m)$ steps, a result comparable to other special architectures laid out for numerical computing, such as certain systolic arrays.[13]

Picture-processing applications

The most interesting application of associative processor systems in the area of array-oriented data organization is picture processing. A digital picture is usually given by an array of pixels (x,y), where x and y denote the coordinates of the pixel within the image. For the dimensions of the two-dimensional pixel array, we shall confine our discussion to a typical square image of N rows and columns. Each pixel has certain gray-level intensities $f(x,y)$ (typically used are 16 gray levels represented by 4 bits or 256 gray levels represented by 8 bits).

The picture is processed at several levels:

- image compression (incoming pixel data is reduced),
- noise reduction and image enhancement (relevant pixels are enhanced),
- feature extraction,
- classification of patterns, and
- analysis of geometric objects.

We can group the used algorithms into three general classes: frequency-oriented algorithms, space-oriented algorithms, and statistical algorithms.[14,15]

Frequency-oriented algorithms consider a picture as a wave pattern—that is, an infinite series of sine and cosine functions represent a picture. The frequencies of these harmonic oscillations give the individual characteristics of the picture.

Space-oriented algorithms interpret a picture as a geometric structure of objects. These objects are characterized by the size, shape, distance to neighbor objects, and the gray-level intensities of their pixels. Space-oriented algorithms comprise operations on local environments of each pixel. Usually, these environments consist of the pixel's four or eight nearest neighbors (see Figure 6).

Statistical algorithms consider a picture as a distribution of gray levels. A histogram usually represents such a discrete distribution. From the distribution, the algorithms derive global transformations of the pixels.

On-line picture-processing architectures comprise peripheral devices from which data enters the system (cameras, integrated sensors); large memory segments (RAM or background devices) to intermediately store the image to be processed; and a processing part to carry out transformations of the picture.

Usually, the data size of an image causes some thread for

READER SERVICE CARD

IEEE MICRO

December 1992 issue (card void after June 1993)

Name _____

Title _____

Company _____

Address _____

City _____ State _____ Zip _____

Country _____ Phone(_____) _____

Reader Interest
(Add comments on the back)

Readers,
Indicate your interest in articles and departments **by circling the appropriate number** (shown on the last page of articles and departments):

150	151	152	180	181	182
153	154	155	183	184	185
156	157	158	186	187	188
159	160	161	189	190	191
162	163	164	192	193	194
165	166	167	195	196	197
168	169	170	198	199	200
171	172	173	201	202	203
174	175	176	204	205	206
177	178	179	207	208	209

Product Information
(Circle the numbers to receive product information)

1	21	41	61	81	101	121	141
2	22	42	62	82	102	122	142
3	23	43	63	83	103	123	143
4	24	44	64	84	104	124	144
5	25	45	65	85	105	125	145
6	26	46	66	86	106	126	146
7	27	47	67	87	107	127	147
8	28	48	68	88	108	128	148
9	29	49	69	89	109	129	149
10	30	50	70	90	110	130	
11	31	51	71	91	111	131	
12	32	52	72	92	112	132	
13	33	53	73	93	113	133	
14	34	54	74	94	114	134	
15	35	55	75	95	115	135	
16	36	56	76	96	116	136	
17	37	57	77	97	117	137	
18	38	58	78	98	118	138	
19	39	59	79	99	119	139	
20	40	60	80	100	120	140	

READER SERVICE CARD

IEEE MICRO

December 1992 issue (card void after June 1993)

Name _____

Title _____

Company _____

Address _____

City _____ State _____ Zip _____

Country _____ Phone(_____) _____

Reader Interest
(Add comments on the back)

Readers,
Indicate your interest in articles and departments **by circling the appropriate number** (shown on the last page of articles and departments):

150	151	152	180	181	182
153	154	155	183	184	185
156	157	158	186	187	188
159	160	161	189	190	191
162	163	164	192	193	194
165	166	167	195	196	197
168	169	170	198	199	200
171	172	173	201	202	203
174	175	176	204	205	206
177	178	179	207	208	209

Product Information
(Circle the numbers to receive product information)

1	21	41	61	81	101	121	141
2	22	42	62	82	102	122	142
3	23	43	63	83	103	123	143
4	24	44	64	84	104	124	144
5	25	45	65	85	105	125	145
6	26	46	66	86	106	126	146
7	27	47	67	87	107	127	147
8	28	48	68	88	108	128	148
9	29	49	69	89	109	129	149
10	30	50	70	90	110	130	
11	31	51	71	91	111	131	
12	32	52	72	92	112	132	
13	33	53	73	93	113	133	
14	34	54	74	94	114	134	
15	35	55	75	95	115	135	
16	36	56	76	96	116	136	
17	37	57	77	97	117	137	
18	38	58	78	98	118	138	
19	39	59	79	99	119	139	
20	40	60	80	100	120	140	

SUBSCRIPTION CARD

SUBSCRIBE TO IEEE MICRO

All the facts about today's chips and systems

❑ **YES**, sign me up!

If you are a member of the Computer Society or any other IEEE society, pay the member rate of only $23 for a year's subscription (six issues).

Society: _____

IEEE membership no: _____

Society members: Subscriptions are annualized. For orders submitted March through August, pay half the full-year rate ($11.50) for three bimonthly issues.

❑ **YES**, sign me up!

If you are a member of ACM, ACS, BCS, IEE (UK), IEEE but not a member of an IEEE society), IECEJ, IPSJ, NSPE, SCS, or other professional society, pay the sister-society rate of only $39 for a year's subscription (six issues).

Organization: _____ Membership no: _____

For information on airmail option, send fax request to (714) 821-4010.

_____
Full Signature Date

_____
Name

_____
Street

_____
City

_____
State/Country ZIP/Postal Code

❑ Payment enclosed *Residents of CA, DC, Canada, and Belgium add applicable sales tax.*

❑ Charge to ❑ Visa ❑ MasterCard ❑ American Express

Charge-card number

_____

Expiration date

____ ____
Month *Year*

Prices valid through 12/31/92
1292 MICRO

Charge orders also taken by phone:
(714) 821-8380 8 a.m. to 5 p.m. Pacific time
Circulation Dept.
10662 Los Vaqueros Circ., PO Box 3014
Los Alamitos, CA 90720-1264

Editorial comments

I liked: _____

_____

_____

I disliked: _____

_____

_____

I would like to see: _____

_____

_____

Reviewers Needed. *If interested, send professional data to Dante Del Corso, Dipartimento di Elettronica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129, Torino, Italy.*

For reader service inquiries, see other side.

PLACE POSTAGE HERE

PO Box is for reader service cards only.

IEEE Micro
PO BOX 16508
NORTH HOLLYWOOD CA 91615-6508
USA

Editorial comments

I liked: _____

_____

_____

I disliked: _____

_____

_____

I would like to see: _____

_____

_____

Reviewers Needed. *If interested, send professional data to Dante Del Corso, Dipartimento di Elettronica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129, Torino, Italy.*

For reader service inquiries, see other side.

PLACE POSTAGE HERE

PO Box is for reader service cards only.

IEEE Micro
PO BOX 16508
NORTH HOLLYWOOD CA 91615-6508
USA

NO POSTAGE NECESSARY IF MAILED IN THE UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO. 38 LOS ALAMITOS, CA

POSTAGE WILL BE PAID BY ADDRESSEE

IEEE COMPUTER SOCIETY
PO BOX 3014
LOS ALAMITOS CA 90720-9804
USA

the data lines to transfer the picture. For example, a digital image of typical size 512 × 512 pixels, each pixel with 256 gray levels, comprises 2 Mbits of data that the data lines must transfer. This causes considerable loading times if the image has to be loaded into memory sequentially in units of, for example, 32 bits. We can now circumvent these data path limitations by transferring optical data directly to memory via arrays of optical sensors integrated into the memory array, as discussed earlier.

Operations on local environments enable massive parallelism because each pixel needs information only from its local neighbor pixels to compute its new value. Local operations can be performed for a large set of pixels simultaneously, if a suitable number of appropriate processing elements is available.

Segments of CAPRA word cells can efficiently carry out these parallel processing tasks: The storage flip-flops of the word cells store the pixel data, and the BOOL units and ALU parts of each word cell work as processing elements.

Input data is scanned into the array through either a row-oriented or a column-oriented method. As most output devices are row-oriented, we shall consider this strategy for the input as well.

Two approaches are possible for memorizing a picture in a CAPRA segment. In the row approach, the image is read in row by row. To process an eight-neighborhood of the pixels of one row for a space-oriented algorithm, it is sufficient to store three rows (the actual one and its upper and lower neighbor rows). For each pixel point of these three rows, r bits in a memory word are used. Figure 7 shows the storage scheme of this approach.

In the so-called block approach, the entire picture is read into a larger CAPRA part. Because an entire row of pixels usually does not fit into one CAPRA word cell, a number of word cells interleavingly store its data, as Figure 8 shows. Each memory word stores $k = n/r$ pixels. The N pixels belonging to one row of the image are stored in the same bit positions of N consecutive memory words, so the corresponding ALUs can analogously process them.

Of course, the block approach necessitates a CAPRA capacity about $N/3$ times as large as that of the row approach. Usually, as an intermediate way between the row approach and the block approach, the CAPRA segment can store a number of $z = w \cdot k / N$ rows of pixels for a given word capacity w of the segment.

These two storage schemes offer to exploit the potential to process the pixels in parallel, as well as to carry out concurrent search operations on them. Thus our architecture appears especially interesting in regard to supporting statistical and space-oriented algorithms.

Many space-oriented algorithms are based on considering each pixel point's corresponding neighborhood. As an example, Figure 9a, next page, shows some approximations for

$$\begin{bmatrix} & f(x,y-1) & \\ f(x-1,y) & f(x,y) & f(x+1,y) \\ & f(x,y+1) & \end{bmatrix} \begin{bmatrix} f(x-1,y-1) & f(x,y-1) & f(x+1,y-1) \\ f(x-1,y) & f(x,y) & f(x+1,y) \\ f(x-1,y+1) & f(x,y+1) & f(x+1,y+1) \end{bmatrix}$$

Figure 6. A local environment of four or eight nearest neighbors for the gray-level value of a pixel point (x, y).

Word cells

$f(i-1,0)$	$f(i,0)$	$f(i+1,0)$
$f(i-1,1)$	$f(i,1)$	$f(i+1,1)$
⋮		
$f(i-1,N-1)$	$f(i,N-1)$	$f(i+1,N-1)$

Figure 7. Scheme of the row approach for memorizing pictures. N consecutive CAPRA words store three rows of the image.

Word cells

$f(0,0)$	$f(1,0)$	•••	$f(k-1,0)$
$f(0,1)$	$f(1,1)$	•••	$f(k-1,1)$
⋮	⋮		
$f(0,N-1)$	$f(1,N-1)$	•••	$f(k-1,N-1)$
$f(k,0)$	$f(k+1,0)$	•••	$f(2k+1,0)$
$f(k,1)$	$f(k+1,1)$	•••	$f(2k-1,1)$
⋮	⋮		
$f(k,N-1)$	$f(k+1,N-1)$	•••	$f(2k-1,N-1)$
⋮			
$f(N-k,0)$	$f(N-k+1,0)$	•••	$f(N-1,0)$
$f(N-k,1)$	$f(N-k+1,1)$	•••	$f(N-1,1)$
⋮	⋮		
$f(N-k,N-1)$	$f(N-k+1,N-1)$	•••	$f(N-1,N-1)$

Figure 8. Scheme of the block approach. The shaded area represents one complete row (row $2k-1$) of the image.

the gray-level gradient using the Prewitt operator, the Sobel operator, and the Roberts gradient. These image operators estimate the value of a given pixel's gray-level gradient using weighted differences of the neighbor points' gray levels.[14,15] Figure 9b shows that matrices of $M=3$ rows and columns also can formally represent the weights.

Because of the different weights, adding up the weighted sum under one instruction stream takes an order of $O(M^2)$ subsequent operations. However, with z rows of a pixel stored in the CAPRA segment, $z \cdot k$ neighborhoods can be processed concurrently. So, compared with the von Neumann architecture, processing speeds up by a factor of $O(z \cdot k)$. Here, spe-

- $\partial f/\partial x\ f(x,y) \approx f(x,y) - f(x+1,y)$

  $\partial f/\partial y\ f(x,y) \approx f(x,y) - f(x,y+1)$

- Prewitt operator

$$\partial f/\partial x\ f(x,y) \approx [f(x+1,y-1) + f(x+1,y) + f(x+1,y+1)] - [f(x-1,y-1) + f(x-1,y) + f(x-1,y+1)]$$

$$\partial f/\partial y\ f(x,y) \approx [f(x-1,y-1) + f(x,y-1) + f(x+1,y-1)] - [f(x-1,y+1) + f(x,y+1) + f(x+1,y+1)]$$

- Roberts gradient

$$\partial f/\partial x\ f(x,y) \approx f(x,y) - f(x+1,y+1)$$

$$\partial f/\partial y\ f(x,y) \approx f(x+1,y) - f(x,y+1)$$

- Sobel operator

$$\partial f/\partial x\ f(x,y) \approx [f(x-1,y+1) + 2 \cdot f(x,y+1) + f(x+1,y+1)] - [f(x-1,y-1) + 2 \cdot f(x,y-1) + f(x+1,y-1)]$$

$$\partial f/\partial y\ f(x,y) \approx [f(x+1,y-1) + 2 \cdot f(x+1,y) + f(x+1,y+1)] - [f(x-1,y-1) + 2 \cdot f(x-1,y) + f(x-y,y+1)]$$

(a)

$$\frac{\partial f}{\partial x} f(x,y): \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix} \quad \frac{\partial f}{\partial y} f(x,y): \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \end{bmatrix}$$

$$\frac{\partial f}{\partial x} f(x,y): \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \quad \frac{\partial f}{\partial y} f(x,y): \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ -1 & -1 & -1 \end{bmatrix}$$

$$\frac{\partial f}{\partial x} f(x,y): \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \quad \frac{\partial f}{\partial y} f(x,y): \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}$$

$$\frac{\partial f}{\partial x} f(x,y): \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \quad \frac{\partial f}{\partial y} f(x,y): \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$

(b)

Figure 9. Approximations for the gray-level gradient at pixel (x,y) of a picture (a), and corresponding 3×3 matrices (b).

cial approximations usually are chosen for the nonexisting neighbor points of the picture's edge points. For instance, these points are assumed to have either some constant, predefined value or to have the same value as the corresponding edge points. Several authors present detailed discussions.[12,14,15]

As one numerical example, the evaluation of the Roberts gradient for an image of 512 × 512 pixels (each with 256 gray levels) takes 0.74 ms using the row approach for a CAPRA segment of 1,024 words of 32 bits. Comparative operations on LUCAS take 3.64 ms, and a conventional VAX computer needs 218 ms.

Whereas the space-oriented algorithms usually operate on very regularly structured subgrids of pixel neighborhood points, statistical algorithms demand evaluation of data properties in very irregularly shaped pixel patterns. The associative operations of the CAPRA architecture very efficiently support the necessary search operations.

To evaluate statistical results, we finally must sum up the hits of these search operations, carried out over the pixel array. This is equivalent to summing up the components of a vector, as discussed earlier. So, for example, in the block approach this phase has a time complexity $O(\sqrt{N^2}) = O(N)$ for a matrix of N^2 pixel elements; on a single processor system it would be of the order $O(N^2)$. Thus, a performance improvement of $O(N)$ results for the CAPRA architecture. Reetz discusses these aspects in more detail.[12]

CAPRA, a new experimental architecture for associative processor systems, comprises several innovative features. They are inclusion of logic elements directly within the word cells or bit cells of memory; use of a maskable decoder to enable multiaccess to the memory and computing devices of the array; activity flags within the cells of the array to enable flexible definition of activity patterns; and integration of sensor elements for the direct parallel input of optical data. This architecture supports database and picture-processing applications. Related application fields like hardware support of neural networks and fuzzy systems already have been investigated by others.[5,13] The CAPRA architecture is an interesting candidate for demonstrating the flexibility and comfort of associative algorithms.

In the future, we plan to complete the hardware realization and to study the described applications, especially under the requirements of real-time systems.

References

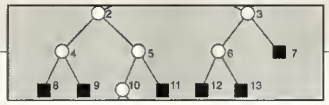
1. K.E. Grosspietsch, "Associative Processors and Memories: A Survey," *IEEE Micro*, Vol. 12, No. 3, June 1992, pp. 12-19.
2. R.M. Lea, "ASP: A Cost-Effective Parallel Microcomputer," *IEEE Micro*, Vol. 8, No. 5, Oct. 1988, pp. 10-29.
3. K.E. Grosspietsch, "Architectures for Testability and Fault Tolerance in Content-Addressable Systems," *IEEE Proc.*, Vol. 136, Part E, No. 5, Sept. 1989, pp. 366-373.
4. D. Tavangarian, "Ortsadressierbarer Assoziativspeicher" ["Location-Addressable Associative Memory"], *Elektronische Rechenanlagen [Electronic Computing Systems]*, Vol. 25, No. 5, Oct. 1985, pp. 264-278.
5. J. Bueddefeld et al., "An Intelligent Sensor Integrated Preprocessing Facility for Neural Networks," *Microprocessing and Microprogramming*, Vol. 32, 1991, pp. 335-342.
6. S.D. Kirkish et al., "Optical Characteristics of CMOS-Fabricated MOSFETs," *IEEE J. Solid-State Circuits*, Vol. 22, No. 2, Apr. 1987, pp. 299-301.
7. H. Onodera, T. Tateishi, and K. Tamaru, "A Cyclic A/D Converter That Does Not Require Ratio-Matched Components," *IEEE J. Solid-State Circuits*, Vol. 23, No. 1, Feb. 1988, pp. 152-158.
8. R. Klinke et al., "Eine Photodetektor-Matrix in Standard-CMOS-Technologie" ["A Photodetector Matrix in Standard CMOS Technology"], *Proc. GME Conf. Mikroelektronik*, VDE/VDI: Gesellschaft Mikroelektronik (GME), Berlin, 1991, pp. 175-180.
9. K.E. Grosspietsch, U. Schaefer, and M. Kohn, "A VHDL Model for the Specification and Simulation of Complex VLSI Architectures," *Proc. Int'l. Workshop Euro-VHDL*, Vol. 276, Swedish Institute of Microelectronics, Kista, Sweden, 1991, pp. 46-52.
10. P. Mishra and M.H. Eich, "Join Processing in Relational Databases," *ACM Computing Surveys*, Vol. 24, No. 1, Mar. 1992, pp. 63-113.
11. C. Fernstrom, I. Kruzela, and B. Svensson, "LUCAS Associative Array Processor," *Lecture Notes in Computer Science*, Vol. 216, Springer-Verlag, Berlin, 1986.
12. R. Reetz, "Parallele Algorithmen und Anwendungen fuer das assoziative Prozessorsystem CAPRA" ["Parallel Algorithms and Applications for the Associative Processor System CAPRA"], *GMD-Studien*, No. 207, Gesellschaft fuer Mathematik und Datenverarbeitung, St. Augustin, Germany, 1992.
13. C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, Reading, Mass., 1980.
14. W.K. Pratt, *Digital Image Processing*, John Wiley & Sons, New York, 1978.
15. R.C. Gonzalez and P. Wintz, *Digital Image Processing*, Addison-Wesley, 1987.

Karl E. Grosspietsch is a researcher at the German National Research Center for Computer Science, Gesellschaft fuer Mathematik und Datenverarbeitung (GMD), in St. Augustin, Germany. He joined GMD in 1974. His main activities comprise research in the fields of computer architecture, fault tolerance, and VLSI design.

Grosspietsch studied computer science at the University of Hamburg, where he received his diploma. He received a PhD in computer science from the University of Bonn. He is a member of the German computing society Gesellschaft fuer Informatik and serves as Associate Editor-in-Chief of *IEEE Micro*.

Ralf Reetz is a PhD candidate in the Department of Computer Science at the University of Karlsruhe, where he is pursuing research in formal hardware verification. He studied computer science at the University of Bonn and received a diploma from there this year.

Direct questions or comments about this article to Karl E. Grosspietsch, German National Research Center for Computer Science, P.O. Box 1316, D-5205 St. Augustin, Germany; or email at grossp@gmdzi.uucp.



A General-Purpose CMOS Associative Processor IC and System

An associative processor architecture integrates the functionality of content-addressable memory, functional memory, and associative parallel processors in a single-chip architecture. We combined a set of 16 such chips to form the Coherent Processor, which interfaces to an IBM PS/2 computer. The processor's writable control store permits quick execution of application-specific microcoded operations.

Charles D. Stormon

Nikos B. Troullinos

Edward M. Saleh

Abhijeet V. Chavan

Mark R. Brule

Coherent Research

John V. Oldfield

Syracuse University

Association has been a long-standing alternative to addressing in computer organization, but its efficient and economic realization has been slow to materialize. Bush originated the concept in his 1945 article,[1] which anticipated many developments we now take for granted. There were flurries of interest in the 1950s and 1960s, as new memory technologies were explored,[2] but cost and, until recently, size have been formidable barriers to effective application. Unlike random-access memory, content-addressable or associative memory is complicated by word length and multiple-hit considerations. Until now, no general-purpose CAM architecture has emerged, although there are specific application niches such as memory management units for fast processors.

We broadly classify associative processors into three types. The simplest and most familiar variety is content-addressable memory. CAM acts as a directory memory: Data stored in CAM is compared in word-parallel fashion to a comparand, which may have certain bits masked. The result is a response vector that indicates whether each word in CAM matched the masked comparand. Typically, the response vector addresses a RAM, which contains the information to be looked up.

The second type of associative processor is functional memory. An FM is similar to CAM except that each cell of the FM can store a third state called "don't care."[3] When a don't-care bit is stored in a cell, it will match either a 1 or a 0 in the comparand. Each word in the FM can store a conjunction of n Boolean variables, where n is the number of bits in each word. Each variable is positional, so if it does not appear in the conjunction, a don't care is stored in that position. The desired output is stored along with the conjunction, to be read out if a match is found. We can form disjunctions by replicating the output at each term in the disjunction. In this way, we can efficiently implement a Boolean function of n variables using the FM, similar to the function of a programmable logic array, except we can selectively change the functions simply by altering the FM contents.

The final variation of associative processing is the associative parallel processor. An APP is a single-instruction, multiple-data (SIMD) parallel computer with a linear interconnection network (bus). Operations are performed on local data based on a CAM-performed selection. To implement an APP, CAM bits of a specified column must be individually writable. We can add an arithmetic logic unit to each CAM word to improve perfomance, but this is not essential.

The APP allows arithmetic, Boolean, and other functions to be performed on the data in CAM, in every selected word in parallel. Both operands may be stored in the CAM word, or one may be in the CAM word and one on the comparand bus. The result is stored in the corresponding

0272-1732/92/1200-0068$03.00 © 1992 IEEE

CAM word. Because of the limitation of the linear interconnection network, the APP cannot perform functions in parallel when the operands are in different CAM words (such as in summation). These must be performed sequentially.

Searching tables and trees

We can easily determine whether an associative solution will be effective in a given application. If a program spends a large proportion of its time searching data structures, managing tables, following index pointers, or performing the same operations on lists or tables of data, an associative solution will probably increase its performance significantly.

CAM has traditionally been used to resolve page table references in virtual memory systems. An associative parallel processor is useful in many other areas. In RAM, we can organize data to optimize one particular access method. In an associative processor, we can organize data to efficiently serve many access methods simultaneously. Consider the problem of storing and operating on binary trees. To store any part of the complete binary tree in CAM, we number the nodes according to the following rules:

- Root node is number 1.
- Left child of node number n is $2*n$.
- Right child of node number n is $2*n+1$.

Following these rules gives a unique number (address) to every node in the tree. When the user program must access the data at a particular node, it accesses the data in one operation by using the address as a search key. A modification to this technique allows more flexible access to the data in the tree. In a fixed-size word, we left-justify the node number, padding to the right with don't cares. For example, Table 1 shows how we would encode the leaf nodes in the tree shown in Figure 1 in an 8-bit field.

The addresses encoded are stored in FM. To access a specific node, we encode the node number in the same manner, left-justifying and padding with don't cares. If the node address is present in CAM, that node will match. Any node that is an ancestor or descendant of the presented node will also respond. For example, for the tree shown in Figure 1, if 2 is presented as the search key (10******), both nodes 4 (100*****) and 5 (101*****) respond. This subtree has node 4 as its root.

Finding the sibling of a particular node is also simplified. Node number 20 is encoded as (10100***). We access its sibling node 21 (10101***) by inverting the least significant bit in the node number. In RAM, this requires either a traversal of the tree or the allocation of more storage to hold pointers to each node's siblings. Graphics and image processing applications use quadtrees (trees with a branching factor of 4) quite frequently. Complex pointer schemes permit efficient access to these trees for a single purpose. By encoding the quadtree nodes (as in the example just discussed), we can perform many operations on the tree equally.

Table 1. Node numbers representing tree nodes.

Node number	Encoded
4	100*****
5	101*****
12	1100****
13	1101****
7	111*****

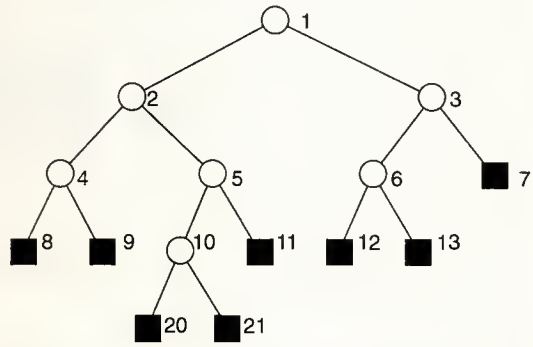


Figure 1. Binary tree.

CAM functions required

The CRC32256 device integrates the functionality of content-addressable memory, functional memory, and an associative parallel processor in a single-chip architecture with 256 processing elements, each having 32 bits of CAM and 4 bits of tag (CAM that is individually bit-writable). We developed the architecture of this device by looking at the requirements of a set of applications and selecting the functions required to support them. The principle of address independence is fundamental to associative processing. Data in the CAM array is accessed without regard to its absolute position in the array. The only position information used is the order that the priority encoder assigns to the words. If a match operation selects a set of words, the encoder provides access to the words in the same order every time.

CAM applications must perform the match operation on a runtime-selectable bit field. Bits in the search argument can be specified as don't cares. Requirements for writing into the memory fall into two categories. The processor must be able to select a subset of words in the memory and alter only those rows. Also, we must provide some number of bits in each word that can be altered without affecting the other bits in those rows.

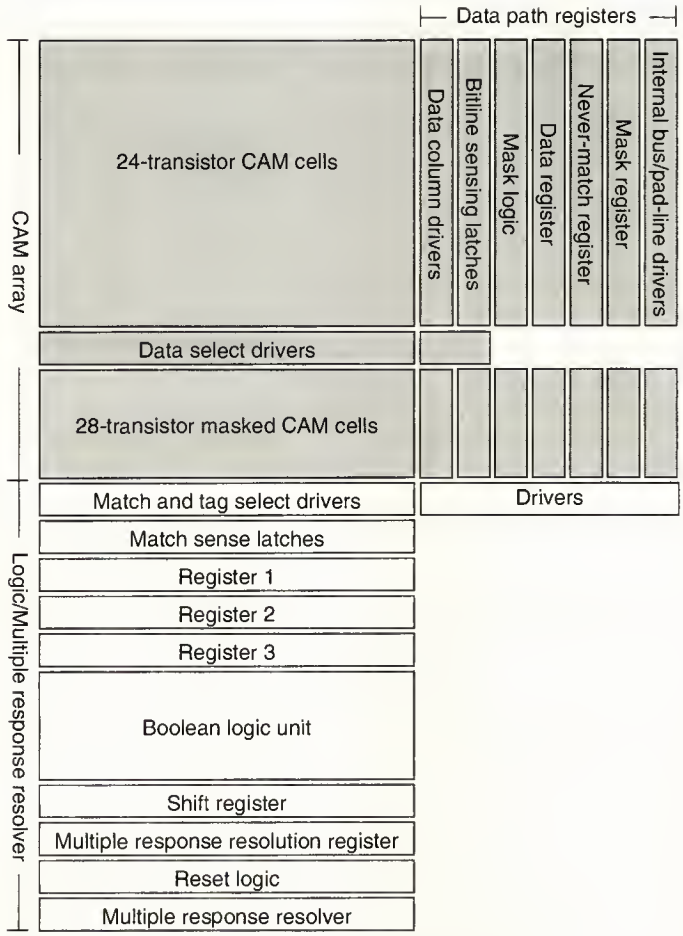


Figure 2. Top-level block diagram of the CRC32256.

The ability to store don't-care values is critical in some applications (such as those involving tree-structured data) but is not needed in all applications. Therefore, the device should store data either with or without don't-care values, and this flexibility should be provided with a minimum impact on storage capacity.

Finally, the device needs a facility for cascading the words horizontally. Some applications fit very nicely into 32-bit words, but many require some other size to store a record of data. Since we intended to integrate the CRC32256 into a system with a 32-bit bus, each memory transaction takes place 32 bits at a time. The user can group multiple 32-bit words together to vary the logical record size.

CAM architecture description

The basis of the CRC32256's VLSI implementation is a modular segment. We used this modular approach partly for the capability to migrate the design to higher densities with more advanced VLSI technologies. The CRC32256 device has a total of two segments, which together form the 256×36 array in a way that is transparent outside the device. As Figure 2 shows, a segment consists of an array of 128 words × 36-bit CAM, the data-path registers, 128 logic rows, a multiple response resolver (MRR) circuit, and all the necessary driving and sensing circuitry. The pitches of the custom-designed cells are all matched in both dimensions. Cascading of the two segments involves four separate mechanisms:

- Enabling for output the data-path drivers of only the active segment during a read operation. A segment is active if it has a word selected.
- Connecting the two halves of the first response register (register 1) through the multiword cascading logic.
- Connecting the two halves of the shift register.
- Completing the MRR tree.

Cascading of multiple devices involves the same mechanisms. When connected in an array, multiple devices form a contiguous area of CAM with associated logic and MRR. The array does not use absolute addressing. On the other hand, adjacency of given words is significant because it provides logical words whose width is a multiple of the physical word width.

The parallel data-path section consists of the data, mask, and "never-match" registers; the read sense amplifier; and the mask-generation logic. All these registers reside on a separate internal data bus that can be selectively coupled to the device data pins. When the mask-enable signal is active, the contents of the data, mask, and never-match registers generate the search argument. This way, the device can store a masking combination, and we can freely intersperse operations that use it with operations that don't, without having to reestablish the masking combination.

Each of the 256 words in the chip has an associated logic row. Two words of CAM are combined and provide three match-result signals and accept two word-select signals. The logic array is organized as a SIMD processor. As Figure 3 shows, each logic row consists of several different registers, switches, and combinational logic blocks. Specifically, a single logic row consists of five registers, three switches, two buffering blocks, two combinational logic blocks, and one Boolean logic unit. We can select any one of the three response registers (registers 1, 2, and 3) to store a match result. The MRR register provides the input to the first stage of the MRR tree. Responses from a match operation must be moved into the MRR register to perform the MRR function on those results. The shift register can be loaded from the MRR register

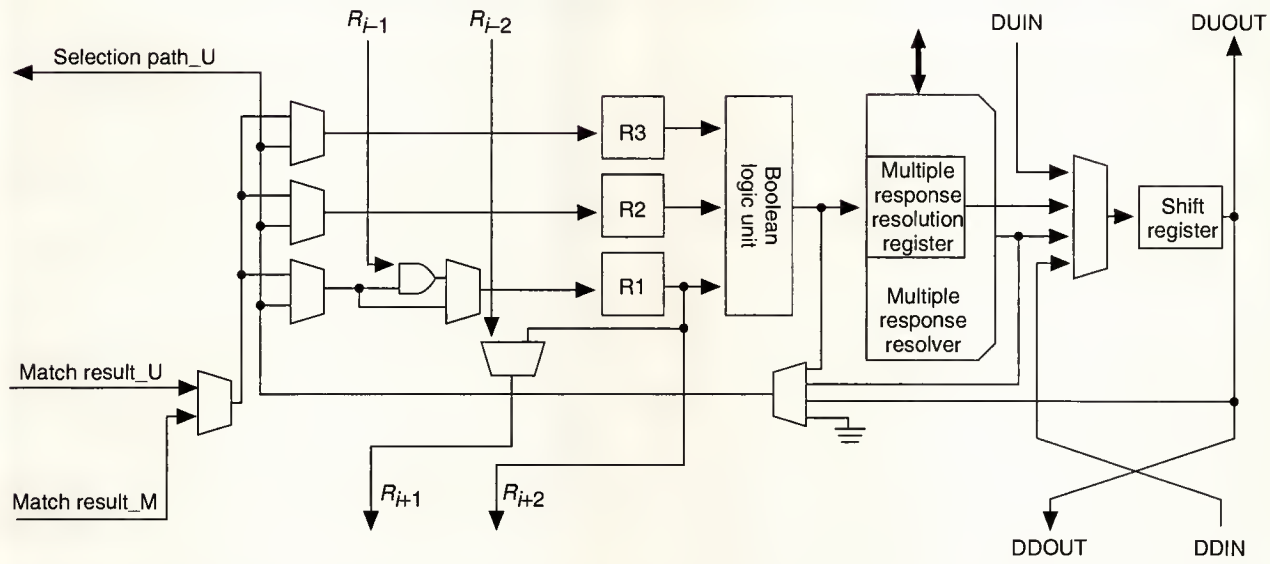


Figure 3. Organization of the logic row.

or from the output of the MRR function. Also, during a shift operation, this register is loaded with the value from the shift register of the word above (shift down) or below (shift up).

Register 1 must be selected as the response register when performing multiple-word matches. As we said, data records can span multiple CAM words. When the application program searches the records, it matches the first word of the comparand against the first word of the CAM records. It then matches subsequent words of the comparand against the corresponding fields within the CAM. As the multiword record matches with a multiword CAM entry, the response in register 1 is propagated through the physical words of the CAM entry. For example, if a data record is four CAM words long, four matches would be performed (one match, followed by three "match-next" operations). Those records that matched would have a response in register 1 that corresponds to the fourth word. The signals R_{i-1}, R_{i-2}, R_{i+1}, and R_{i+2} provide the necessary communication between adjacent CAM words for multiword response capability. Bit-mode operations use R_{i-1} and R_{i+1}, and quad-mode operations use all four signals.

The MRR function allows the topmost response in the CAM array to be selected for reading or writing. The output of the MRR register from a given row will be 1 if there is a response in that row. Through the MRR function, these outputs combine to generate a response signal for the segment, and then for the entire device. When the device is accessed for reading or writing through the MRR, the enable signal first passes to only one segment, depending on where the topmost response is in the linear array. This enable signal then passes back through the MRR tree of that segment in a mutually exclusive fashion, until it ultimately arrives at the topmost row of the segment. This topmost row will be the only row in the array to generate an MRR output signal, and thus it will be the only row accessed during the read or write operation.

Further, when a "select-next" operation executes, the row that has MRR output set will reset the contents of its MRR register to 0. Now when the MRR function is evaluated, the enable signal will be sent to the next row that has a response. If this row is in the next segment (or device, for an array of chips), the cascading logic transparently steers the enable signal. All this happens in a time proportional to the logarithm of the array size, which for reasonable array sizes is one clock period.

The Boolean logic unit can perform one of 256 logic functions to registers 1, 2, and 3, and their inverted outputs. The Boolean logic unit output can be stored in the MRR register or registers 1, 2, or 3; or it can be used as the select path signal. This block provides the interconnection path for transferring data between registers, as well as the means to modify these registers' contents. In addition, the use of the Boolean logic unit output as the select path signal allows access to a CAM word based on any combination of the response registers.

The select path is the signal that accesses the CAM word. This signal is enabled when a CAM word is read or written. A word can be selected via one of four paths: MRR output, Boolean logic unit output, the shift register, or an unconditional selection of all words.

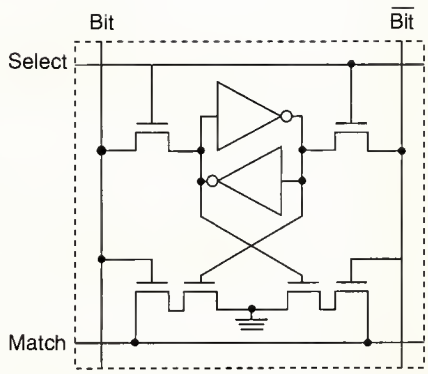


Figure 4. Static CAM cell.

Table 2. Encoding scheme for quad values in CAM.

1	0	Don't care	Never match
11	00	10	01

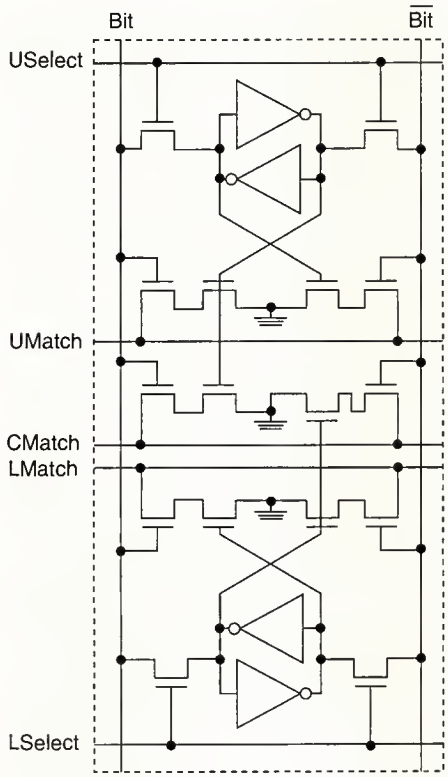


Figure 5. Static CAM cell for quad word storage.

Circuit specifics

In the CAM array design, our first concerns were simplicity and performance. We extensively simulated and verified the CAM cell, the basic logic row, and the MRR block in small test chips. The basis of the CAM cell design is the 10-transistor static cell shown in Figure 4. We verified this building block in a test structure of 32×32 cells with driving, sensing, and scan-path test circuitry.

To store don't-care and never-match states, we combined two CAM cells vertically with a third match line and associated comparator. (Matching against a don't-care state will always yield a response, whereas matching against a never-match state will never yield a response.) The don't-care state presents a double 0 to the inputs of the middle match-line comparator; the never-match state presents a double 1 to these inputs (Table 2). Figure 5 shows this extended CAM cell architecture. The four transistors between the two 10-transistor cells implement the comparator for the middle match line. A row of these new 24-transistor cells provides storage for two binary words or for a single quad word, depending on which match line is used.

The 24-transistor twin cell is 44×98 microns, using 2-micron MOSIS rules. It is 20 percent taller than a pair of 10-transistor cells stacked vertically. By providing the capability to use the upper and lower halves separately when operating in a binary mode, we maintained a good balance between functionality and chip area. Also, the decision to view the storage area as consisting of bits or quads alternatively made the overall chip design simpler: We passed the details of quad reading and writing to the user. The tag bits provide bitwise selective writing. A vertical select line in each column controls two extra select transistors.

The CAM array's parallel nature presents special difficulties in achieving robust write and match operations.[4] With multiple write, the pull-ups in each cell present a resistive load to the column drivers. The strength of the pull-ups, the vertical size of the array, and the strength of the drivers have to be carefully balanced, given power limitations and acceptable noise levels. For match operations, all match lines must be precharged, and then single-rail sensing must occur in the presence of noise from the search argument broadcasting drivers. These energy-intensive operations have no counterparts in static RAM design, and we had to respect the dynamic power limits of a CMOS device.

We used Mead and Wawrzynek's complementary set-reset logic (CSRL) design principle[5] extensively in the on-chip registers. The data-path registers—including bit-line sense amplifier, match sense amplifiers, response registers, and shift register circuits—are variations of the basic CSRL design. We used CSRL because it combines static retention, fast-sensing action, and low power in one compact circuit.

A variation, which we believe is original, is the implementation of a single-phase clock master-slave structure using

dual CSRL latches cascaded in series (a modified N stage followed by its dual P stage). Figure 6 shows this version of CSRL, where the first stage, the master, is a memory cell with differential inputs In and $\overline{\text{In}}$ that sense when the clock input is high and hold when it is low. The second stage, the slave, has the output and complementary output of the master as its differential inputs. This stage senses when the clock input is low and holds when it is high. The outputs of the slave stage, Q and $\overline{\text{Q}}$, are the actual outputs of the register. Mead and Wawrzynek provide further details on CSRL.[5]

The MRR is one of the more complex circuits in the chip (see Figure 7). When a search of the CAM array results in more than one match, the MRR allows these matching entries to be accessed in a prioritized fashion. The MRR circuit has a tree structure to ensure fast operation. Its speed is critical, because this circuit must combine results from all rows in the array (feedforward path) and then steer the enable signal to the topmost row (feedback path). To achieve maximum speed while using minimum space, we used pseudo domino logic in this block. The feedforward logic uses modified domino logic,[6] while the feedback logic uses full domino logic.[7] It is very important to avoid charge redistribution problems when adopting this combination of logic families.

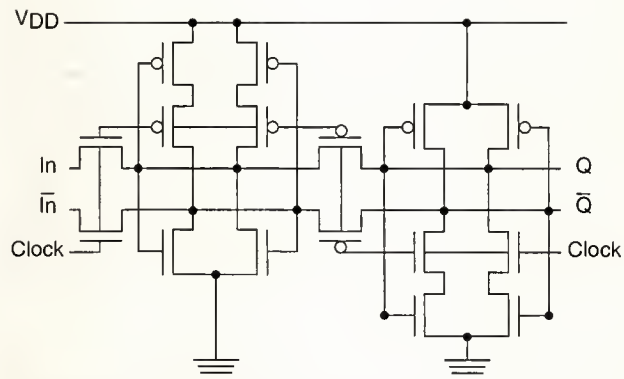


Figure 6. Master-slave CSRL latch with one clock.

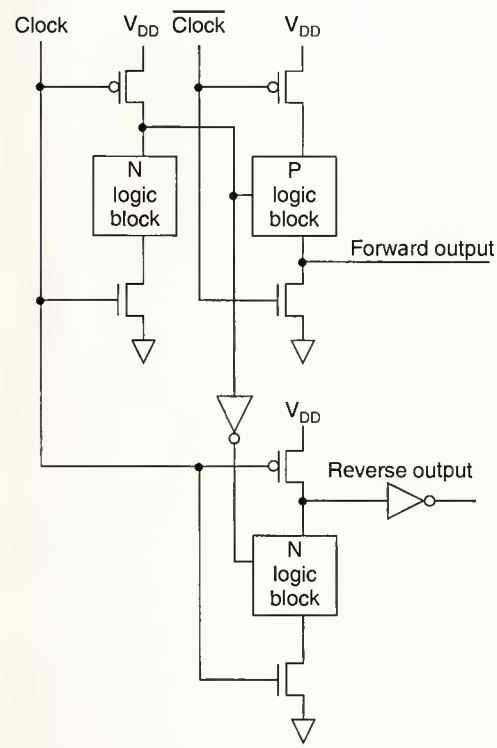


Figure 7. Organization of a single MRR block.

Chip performance

The complete chip in 2-micron CMOS combines 256 rows, each with 36 bits of CAM, and the row logic pipeline. Control is largely external to allow overlapping of control sequences and broadcast of these sequences to multiple chips. In the worst case, control signals and data must be valid for at least 20 ns, which translates to a maximum microcode clock rate of 50 MHz. Complete operations take from two to five control words and may be partially overlapped. Table 3 (next page) gives typical operation times. The CSRL sensing latches used as sense amplifiers typically exhibit a 15-ns sensing time for the single-ended match lines and 8 ns for the complementary bit lines.

The 9.2×9.2-mm die size includes pads. The chip contains approximately 200,000 transistors and is packaged in a 108-pin ceramic pin-grid array. As with all full CMOS devices, the power dissipation depends only on the rate of switching. Despite the high energy requirements of the parallel operations, the average power consumption remains below 400 mW, even at the maximum clocking rate.

System hardware design and environment

Since the CRC32256 is a microcode-controlled device, it can work with a variety of system architectures. Design possibilities range from tightly coupled memory subsystems to loosely coupled processing subsystems. The first such integration of the CRC32256 is a microchannel memory device, known as the Coherent Processor.

We designed the Coherent Processor to resemble system memory; thus the interface is a 32-bit slave type. Accesses to the Coherent Processor are made through function calls, which translate to memory-mapped reads and writes. These memory accesses are decoded to control a microcode sequencer, which passes prestored control sequences to the CRC32256 pins. The Coherent Processor has an array of 16 CRC32256 components, totalling 4,096 associative processing elements. Re-

Table 3. Operation times.

Operation	Clocks	Maximum time (ns)
CAM write	1	20
CAM match	2	40
CAM read	3	60
Shift	2	40
Select next	4	80

turned values from read functions are either contents from a specific CAM word or flags signifying the Coherent Processor array's status. Write functions provide a 32-bit input value, which is stored in a specific CAM word, matched against the CAM, or used as control data. Runtime access to the Coherent Processor causes the microcode sequencer to present control words to the array until it encounters a stop bit in the microcode. At the end of a sequence, control returns to the host CPU, along with any requested return value.

Four Xilinx field-programmable gate arrays provide decoding, registers, and alternative data and control paths. Under the runtime environment, the microcode sequencer controls presentation of the microcode to the array. Using the single-step function, it may present microcode one word at a time from the system bus. This lets us scrutinize the performance of the CRC32256 in a multichip configuration. The microcode presented to the array by the sequencing engine is retained in a writable control store. High-speed (25-ns) static RAM modules implement the 16,000×64-word store. Its size permits many sequences, and thus it can accommodate a broad range of application-specific macros for the Coherent Processor. The user program loads appropriate microcode into the writable control store during initialization of the Coherent Processor. Providing the correct offset address into the store executes the individual sequences.

System software design and environment

To make application development easy, we designed the Coherent Processor software environment to be simple, understandable, and powerful. The Coherent Processor development system, shown in Figure 8, consists of an assembler and linker for writing programs, as well as a software simulator and source-level debugger. The assembler reads files containing statements in the Coherent Processor macro assembly language. The language specifies memory and logic operations using an assignment syntax. For example, the instruction

```
R2 = Match(0,1000,Mask,Bit)
```

specifies a match operation matching the 32-bit value 0 with the contents of the data part of the CAM array, and the binary value 1000 with the four tag bits. It also specifies that masking should be enabled, that the match is performed in bit mode, and that the results are to be stored in register 2.

The assembler lets the programmer write custom Coherent Processor operations. For example, the following instruction specifies that the contents of the first row, where both registers 1 and 2 contain a 1, are to be read:

```
GetR1AndR2:
        MRReg = R1 ^ R2,
        *0 = CAM[MRROut].
```

This GetR1AndR2 operation stores the logical AND of registers 1 and 2 in the MRR register. Next, multiple responses are resolved, and the word of memory indicated by the topmost responder (CAM[MRROut]) is read out. The result is placed in *0, which indicates that it should be placed in position 0 in a user data structure. This operation is called from a C program by a procedure named callCP, as follows:

```
callCP(GetR1AndR2,mydata).
```

These callCP procedure calls are placed in a C program compiled and linked to a special Coherent Processor library. The result of the read operation is placed into mydata[0], where we assume that mydata is an array name. The programmer specifies the index (0) in the macro assembly code. The assembler's output is a file included in the user's C program and contains a data array defining the contents of the writable control store. A set of definitions lets the code reference operation names (for example, GetR1AndR2).

Applications

In many applications, the Coherent Processor can provide a cost-effective performance increase. Traditional applications of CAM have been in virtual-memory-translation look-aside buffers and local area network routers. The Coherent Processor's increased functionality opens up a host of additional applications.

CAM application. When searching text, we frequently need to find a short string within a longer string. This operation is called a substring search. The Coherent Processor can store text by putting four 8-bit ASCII characters into each 32-bit CAM word. To find any occurrence of a substring, we present that substring as a search pattern four times, because the substring could begin in any of the four characters stored in a word. For example:

```
match ("A B C D")
match ("* A B C")
matchnext ("D * * *")
match ("* * A B")
```

```
matchnext ("C D * *")
match ("* * * A")
matchnext ("B C D *")
```

Using this algorithm, the substring search takes the same number of operations as the number of characters in the short string. The search is independent of the long string's size, provided that the long string fits in CAM. We can extend this method to situations where the search string is "fuzzy." We use don't cares for parts of the search string that are not completely specified and perform multiple searches.

FM applications. Several functional memory applications experience increased performance.

Quadtree Manhattan rectangle generation. In quadtree software, it is usual to apply a recursive divide-and-conquer algorithm to generate the quadrants for a given Manhattan rectangle. But if we use a quadtree variant of the scheme illustrated for Figure 1—that is, with four children per parent, and storing the quadtree leaves in FM—the order is immaterial. Thus we can generate the quadrants directly with a "covering sequence" method,[8] which covers the x- and y-coordinate ranges by minimum sequences and then forms their Cartesian products. For example, the integer range 3 through 9 is covered by the sequence 0011, 01**, and 100*. This method is much faster, particularly for small rectangles, and also permits nonstrict quadrants whose sides are $2^{**}m \times 2^{**}n$.

Region growing. Image processing often requires grouping together contiguous regions of pixels with the same or a similar color. The algorithm for grouping is called region growing, and it can benefit greatly from the use of functional memory.[9,10]

The first step is to encode the image with a quadtree representation.[8] Next, we search the image for the quadrants with the desired color and store the set of responses in register 2. Picking one of those quadrants, we label it region 1 and remove this quadrant from the list in register 2. Picking the next quadrant from register 2, we see if it is a nearest neighbor (in image space) of the region 1 quadrant (four CAM searches). If it is, we mark it region 1; otherwise, we mark it region 2 and remove it from the list in register 2, and so on.

Generally, for each quadrant that has the correct color, we search for all neighboring quadrants of the same color. If there is none, we name the current quadrant a new region. If there is at least one match, but none already labeled, we name them all a new region. If there is at least one match and at least one is already labeled, we rename the current quadrant to the region named by the first responder, and rename all connected regions to that same region name.

The remaining set of regions represents the set of contiguous coherent areas in the image. The computation time is proportional to the number of quadrants that have the correct color.

Pattern and symbol recognition. This application compares a library of patterns against an incoming pattern. Each library element is constructed over a training period by overlaying two or more example patterns and making them don't-care bits where they differ. Thus, each library element is a template with which to match incoming patterns. By carefully selecting the categories and the examples to be combined, we can use each library pattern to recognize a class of input patterns. A symbol is a special case of a pattern, which is built of pixels that represent a symbol to a human observer.

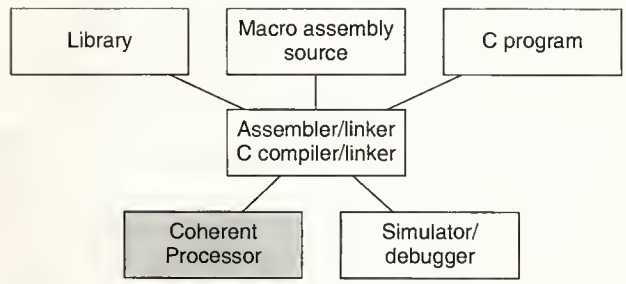


Figure 8. Coherent Processor development system.

With the library of templates stored in FM, we can quickly categorize an incoming pattern. The incoming pattern is the comparand, and any template that matches it indicates the pattern category to which it belongs. The trouble with this method is that it is very sensitive to noise and distortion (rotation, scale, translation, and so on).

We can easily overcome scale and translation distortion by preprocessing the pattern. Correct rotation can often be deduced by contextual clues. If not, either applying a rotation-invariant transform (for example, the Hough transform) or presenting the pattern in a variety of rotations should be effective. We can eliminate random noise by applying the region-growing algorithm just reviewed and eliminating any regions less than a certain size. The quadtree representation will also reduce the amount of FM needed for patterns with a good deal of coherency. Provided that the library fits in the FM, the pattern-recognition step requires only a single associative search (not including whatever preprocessing is necessary).

Prolog accelerator. An experimental Prolog compiler for the CP[11] uses the CAM to provide parallel backtracking and to improve the efficiency of the well-known Warren Abstract Machine (WAM).[12] The compiler stores instantiations (values bound to a variable) for variables in CAM and retrieves them by presenting the variable name as the search key. This approach eliminates the overhead of creating space for unbound variables, dereferencing argument registers for these variables, and trailing the various bindings of one variable to another. In this model, the CAM stores the Prolog terms and the global stack. An abstract instruction set, based on this CAM storage model, is similar to but simpler than the WAM, simplifying

both the compiler and the runtime system. In addition, operations on lists, garbage collection, and the occur check are faster and more efficient because we store variables and terms in CAM.

An earlier development of an experimental Prolog interpreter explored how other aspects of Prolog execution can be accelerated using CAM.[13] In addition to the stack management used in the compiler, the interpreter also explores clause filtering, and unification of lists and other data structures.

Clause filtering is an indexing technique that uses CAM to store a superimposed code word for the head of each clause in the Prolog program. When the current goal is to be executed, the interpreter computes and presents its code word to the CAM as a search key. Only clauses that match the functor, arity, and any instantiated variables are candidates for unification. This technique reduced the number of unnecessary unifications by a factor of 2 to 20, depending on the program run.

By representing list structures in CAM with an efficient tree representation, we can also reduce the time to unify one structure with another. The Prolog interpreter's unification algorithm reduces the complexity of comparing lists to, at worst, the number of variables and ground terms in the larger list. Conventional structure unification requires the algorithm to traverse each list sequentially. The more complex the nested list structure, the better the CAM algorithm performs by comparison.

Parameter window addressable memory. We can implement a broad set of applications efficiently using a functional memory combined with the covering-sequence algorithm (see the section on the quadtree Manhattan rectangle algorithm) for computing the set of 1,0,* patterns required to represent a range of integers. A parameter window is a region in an n-dimensional space represented by n parameters and the set of values these parameters may assume within the window. If we restrict the set of values to a single contiguous range of values, the parameter window will be rectangular. We can represent more complex parameter windows by combining a set of rectangular windows. We represent a parameter window in functional memory by finding the set of rectangles that covers the window, finding a covering sequence for each parameter for each rectangle, forming the Cartesian product of the covering sequences, appending the name of the window, and storing the resulting words in FM with each parameter in a corresponding positional field. Given that, finding whether a point is in one or more parameter windows requires a single search.

Determining which parameter windows cross another selected window amounts to computing the covering sequence and presenting it as a sequence of comparands. The search time is independent of the number of windows—a very useful property because this problem tends to be exponential in the number of dimensions when performed sequentially.

Rule-base accelerator. Kogge et al.[14] have experimented with CAM support for production systems and found it to provide up to two orders of magnitude performance improvement over conventional approaches. OPS5, a forward-chaining inference system, scans facts to determine which rules' conditions are satisfied. From this set, OPS5 selects a single rule and modifies the facts according to that rule. The cycle then repeats.

By compiling OPS5 programs into a structure known as a Rete net, conventional implementations make this comparison as efficient as possible. Using a CAM to store the components of the Rete net further increases the efficiency because a single CAM search can identify all rules whose conditions may be satisfied by a given fact, or conversely all facts that fit the conditions of a given rule. In addition, since facts in OPS5 frequently need to be created or destroyed, the functional memory can identify all instances of a fact and free them immediately rather than traverse the network looking for instances of a fact that needs to be deleted.

APP applications. Two applications are particularly interesting.

Neural network simulation. When processing the inputs to each layer of a neural network, the system multiplies each input by a weight specific to that connection, sums the weighted inputs to each neuron, and compares them to a threshold. If we represent each connection and its associated weight in the Coherent Processor, the system can fan out each input as needed by multiplying all the connection weights in parallel.

The summing and thresholding steps are combined so that we check whether any weighted input alone is enough to exceed the threshold. If not, we subtract the largest weighted input from the threshold and try again. We continue until the neuron fires or we exhaust the nonzero weighted inputs. This is usually much faster than computing the summation, particularly because the Coherent Processor performs the searches in parallel.

Sparse matrix computations. Just as in the neural network application, we do not need to spend any time on matrix terms whose values are zero or some other default value. The Coherent Processor's CAM stores the nonzero entries in the matrix, along with their row and column numbers. The zero entries are simply not stored. Now we can perform any operation on a row or column in parallel. Likewise, we can arrange particular patterns of parallel processing (such as multiply every other entry in every other row by 2).

We move data in the matrix by rewriting the row and column identifiers to appropriate new values. Naturally, searching the data becomes a unit operation. This technique is useful in Gaussian elimination, Fourier transforms, discrete set operations, and virtually any other matrix operation in which the data is sparse—because no time is spent on zero entries.

WE ARE WORKING ON SEVERAL IMPROVEMENTS in the system design. First, bit-selective update and the storage of don't cares are so essential that we will optimize future designs for these capabilities. A promising approach would be to combine a complementary dynamic CAM cell like that of Sodini and Wade[15] (in which the storage of don't cares comes naturally because of the lack of feedback) with a two-dimensional selection mechanism.

We can easily achieve higher density through smaller geometry fabrication techniques. Scaling for smaller geometries is straightforward, given the hierarchical design. At 0.8 micron, a device of 1,000 words (eight segments, 36 Kbits total) would require a die size of 7×7 mm. By generating more of the control internally (and trading some performance), we can reduce the pin count to 68.

We are also designing a second-generation Coherent Processor board-level architecture to make the Coherent Processor more autonomous. Adding direct memory access capability is one way we will achieve this goal. Also, we want to decouple the execution sequences from the host processor using a scheme in which the host CPU passes a command to the Coherent Processor, and the Coherent Processor signals the host when it has completed a function. Together these efforts will result in an efficient single-board system with 64,000 processing elements and 256,000 bytes of CAM.

Acknowledgments

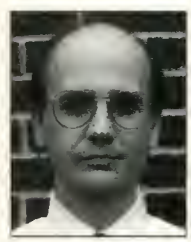
We thank the Syracuse University CASE Center, the IBM Federal Systems Division, and the USC/ISI MOSIS Service. Particular thanks are due to P.D. Kogge, J.A. Coleman, J.C.R. Ribeiro, K.A. Greene, G.C. Stiles, D. Elmendorf, J.A. Robinson, R.D. Williams, N.E. Wiseman, J.D. Kim, K.E. Twardowski, K. Mcveary, T. Park, S. Ramirez, and H. Krempel.

References

1. V. Bush, "As We May Think," *Atlantic Monthly*, Vol. 176, July 1945, pp. 101-108.
2. T. Kohonen, *Content-Addressable Memories*, 2nd ed., Springer-Verlag, Berlin, 1987.
3. T. Kohonen, *Associative Memory*, Springer-Verlag, Berlin, 1977.
4. N. Troullinos and C. Storman, "Design Issues in Static Content-Addressable Memory," Case Center Tech. Report 9208, Syracuse, Univ., Syracuse, N.Y., Aug. 1991.
5. C. Mead and J. Wawrzynek, "A New Discipline for CMOS Design: An Architecture for Sound Synthesis," *Proc. Chapel Hill Conf. Very Large Scale Integration*, Computer Science Press, Rockville, Md., 1985, pp. 87-94.
6. V. Friedman and S. Liu, "Dynamic Logic CMOS Circuits," *IEEE J. Solid-State Circuits*, Vol. SC-19, No. 2, Apr. 1984, pp. 263-266.
7. R.H. Krambeck, C.M. Lee, and H.F. Law, "High Speed Compact Circuits with CMOS," *IEEE J. Solid-State Circuits*, June 1982.
8. J.V. Oldfield et al., "Content-Addressable Memories for Storing and Processing Recursively Subdivided Images and Trees," *Electronics Letters*, Vol. 23, No. 6, 1987, pp. 262-263.
9. R.D. Williams, *Quadtree Operations for Content-Addressable Memories*, doctoral dissertation, Univ. of Cambridge, Cambridge, UK, 1988.
10. W.E. Snyder et al., "Content-Addressable Read/Write Memories for Image Analysis," *IEEE Trans. Computers*, Vol. C-31, No. 10, Oct. 1982.
11. H. Bacha, "Beyond the WAM, a PAM for the CAM," Case Center Tech. Report 8816, Syracuse Univ., Syracuse, N.Y., Dec. 1988.
12. D.H.D. Warren, "An Abstract Prolog Instruction Set," Tech. Note 309, SRI Int'l, Menlo Park, Calif., 1983.
13. C.D. Stormon et al., "An Architecture Based on Content Addressable Memory for the Rapid Execution of Prolog," *Proc. Fifth Int'l Conf. and Symp. Logic Programming*, 1988, pp. 1448-1473.
14. P.M. Kogge et al., "VLSI and Rule-Based Systems," in *VLSI for Artificial Intelligence*, J.B. Delgado-Frias and W.R. Moore, eds., Kluwer, Boston, 1989, pp. 95-108.
15. C.G. Sodini and J.P. Wade, "Dynamic Cross-Coupled Bitline Content Addressable Memory Cell for High Density Arrays," *IEEE J. Solid-State Circuits*, Vol. SC-22, Feb. 1987, pp. 119-121.

Charles D. Stormon is CEO and chief scientist of Coherent Research. While pursuing a PhD in computer engineering, he founded the company in 1987 with the assistance of the New York State Center for Advanced Technology in Computer Applications and Software Engineering (the CASE Center) at Syracuse University. His technical interests lie in computer systems engineering, neural networks, parallel processing, and knowledge-based systems.

Stormon received his BS and MS degrees in computer engineering from Syracuse University. He has won multiple Small Business Innovation Research (SBIR) awards. His affiliations include an adjunct faculty position at Syracuse University's College of Engineering, membership in the IEEE and the Computer Society, board membership in the Northeast Parallel Architectures Center and the Technology Development Organization of Central New York, and a manuscript reviewer's position for *IEEE Transactions on Computers*.



Nikos B. Troullinos is a senior design engineer at Coherent Research, working on new-generation CAM devices. He was involved with the development of the associative processing architecture presented in this article, focusing on circuit and physical design. He is also a PhD candidate in

the Department of Computer and Information Science at Syracuse University, where he works on reduction schemes for the lambda calculus as they apply to architectures targeted to declarative languages.

Troullinos received his electrical engineer's diploma from the Aristotle University of Thessaloniki, Greece, and his MS degree in computer engineering from Syracuse University. He was awarded a Fulbright Fellowship. He is a member of the IEEE Computer Society and the ACM.

Edward M. Saleh is a computer engineer at Coherent Research involved in the design of associative processor VLSI and system architectures. He has worked as a computer engineer for General Electric, where he participated in the design and development of the Warp Systolic Array Processor. His professional interests include parallel processing computer architectures, specifically SIMD machines. His master's work was on fault tolerance in parallel processing architectures.

Saleh received his BS in electrical engineering from the State University of New York at Buffalo and his MS in computer engineering from Syracuse University. He is also a graduate of the General Electric Edison Engineering Program and the General Electric Advanced Course in Computers.

Abhijeet V. Chavan is an advanced project engineer with the Delco Electronics division of General Motors. When this article was written, he was a design engineer with Coherent Research. He has also worked for Siemens AG as a design engineer for industrial automation systems and as a research assistant at the CASE Center at Syracuse University. His research interests include systems engineering, VLSI design, VHDL, and object-oriented programming.

Chavan received his BS degree in electrical engineering in 1983 and his MS in computer engineering in 1988.

Mark R. Brule is vice president of engineering at Coherent Research. He currently works with image processing and pattern recognition applications using CAM technologies. Previously, he worked for Genigraphics as a software engineer. His master's work concerned approaches to logic programming execution using content-addressable memory.

Brule received his BS degree in computer science and his MS degree in systems and information science from Syracuse University. He is currently finishing his MS in mathematics at Syracuse University.

John V. Oldfield, a professor in the Department of Electrical and Computer Engineering at Syracuse University, presently is on leave as a visiting professor at Tektronix. In 1983 he started the Syracuse University Machines for Associative Computation (SUMAC) project, under the aegis of the CASE Center at Syracuse University. He has held academic and research positions at Bruce Peebles, Ltd., Edinburgh; the University of Edinburgh; University College Swansea, and the University of Wales. His research interests include content-addressable memories, with applications in computer graphics and data and image compression; field-programmable gate array architectures; and self-timed systems.

Oldfield received his BSc and PhD degrees from the University of London. He is a senior member of the IEEE and a member of the IEEE Computer Society.

Direct questions concerning this article to Edward M. Saleh, Coherent Research, Inc., 1 Adler Drive, East Syracuse, NY 13057, or via e-mail at saleh@cri.com.

1992 INDEX
IEEE MICRO

This index covers all technical items—papers, correspondence, reviews, etc.—that appeared in this periodical during 1992, and items from previous years that were commented upon or corrected in 1992.

The Author Index contains the primary entry for each item, listed under the first author's name, and cross-references from all coauthors. The Subject Index contains several entries for each item under appropriate subject headings, and subject cross-references.

It is always necessary to refer to the primary entry in the Author Index for the exact title, coauthors, and comments/corrections.

AUTHOR INDEX

SUBJECT INDEX

PC miscellany

Richard Mateosian

2919 Forest Avenue

Berkeley, CA

94705-1310

(510) 540-7745

This time I've focused on PC-related books. Some are better than others, but the general trend of books in this area is upward.

Borland Books explain Microsoft Word

Microsoft Word for Windows 2.0 is an enormously complex product. Its *User's Guide* contains 860 pages of fine print. Neither the *User's Guide* nor Word's chaotic user interface gives users much help in using the power and flexibility of this word processor. Two books from Microsoft Press, both by Russell Borland, help fill this gap.

Running Word for Windows, Version 2, Russell Borland (Microsoft Press, Redmond, Wash., 1991, 585 pp.; $27.95)

Borland provides what Microsoft failed to provide in its product offering—a visible, coherent structure. Early in the book he introduces the four pillars of Word for Windows: styles, fields, macros, and templates.

I am not reviewing Microsoft Word for Windows here, but I just reread my December 1987 review of Word 3.01 for the Macintosh. It amazes me how completely I take for granted now the features I raved about then. The problems with Word I had then centered around styles and macros. Borland's book makes clear how much this situation has improved.

I don't want to go into the details of these features now, but if you want to use Word for Windows, you should get this book and read it. The Microsoft *User's Guide* is still the ultimate reference, but Borland's book provides a much-needed missing element—order out of the chaos.

Microsoft Word for Windows 2.0 Macros, Russell Borland (Microsoft Press, Redmond, Wash., 1992, 487 pp. and a 3.5-inch diskette; $34.95)

Microsoft started, as everyone knows, with Bill Gates running around to computer hobbyist meetings giving away paper tapes of his Basic interpreter for 8080-based machines. Basic is an acronym for Beginner's All-purpose Symbolic Instruction Code. Like Pascal, it is a teaching language. Its designers, Kemeny and Kurtz of Dartmouth, never intended it to be used for "real programming," but many people used it that way in the 1970s, before the C language achieved its current widespread popularity.

Now Microsoft has found an ideal use for Basic. They have used it as the basis for an excellent and powerful macro facility for Word for Windows. Word Basic gives users access to all of Word's built-in functions. Anything you can do from a menu and a dialog box, you can do from Word Basic. In fact, Word provides a facility for monitoring user actions and translating them into Word Basic commands.

Borland discusses macros in *Running Word for Windows*, but proper treatment of this facility really requires a book of its own. Like *Running Word for Windows*, this book is tutorial in tone and is organized around topics. However, its appendixes, comprising nearly half the book, provide a technical reference for Word Basic. Since Microsoft's *User's Guide* covers Word Basic sketchily, Borland's book is essential for anyone who needs to use Word macros seriously.

Scheherezade as hacker

Jamsa's 1001 DOS and PC Tips, Kris Jamsa (Osborne/McGraw Hill, Berkeley, Calif., 1992, 896 pp. and a 3.5-inch diskette; $39.95)

This book contains 1,001 consecutively numbered suggestions on how to use your PC and

0272-1732/92/1200-0086$03.00 © 1992 IEEE

DOS more effectively. The pages are unnumbered; the table of contents and the index refer exclusively to tip number.

The book is well integrated with the companion diskette. A special icon ties the book and diskette together. The icon, a dog with a diskette in its mouth (a computer user's best friend), appears in the margin with a file name under it. The current tip refers to the named file on the diskette, telling you how to use it to implement the suggestion Jamsa is giving you. This is a big improvement over the usual case, in which the companion diskette seems like an afterthought, culled from the book after it was written.

Of course, the value of the book depends on the quality of the information. As I paged through the book, I found many fascinating facts and suggestions. On the other hand, as I looked closely, I quickly found examples of careless writing and editing. For example, Jamsa directs you to type the command DEBUG<LOCKOUT.SY, when he clearly means DEBUG< LOCKOUT.SCR to create the file LOCKOUT.SYS. On the same page he has misspelled the word "different." A misspelling is merely annoying, but feeding the wrong file to the DEBUG program can have catastrophic results.

Another confusing point occurs shortly afterward in tip 5. He says that his illustration shows the default scan lines used by monochrome (scan lines 6 and 7) and color (scan lines 11 and 12) video displays. In fact, the illustration shows two 8-by-8 arrays of empty boxes. Neither array shows which boxes correspond to a cursor. The monochrome array has lines numbered from 0 to 7. The color array numbers the same lines with pairs of hexadecimal digits, from 0-1 to E-F. There are no lines 11 and 12. The hexadecimal codes corresponding to 11 and 12 are B and C, which don't correspond to anything in the picture.

As all authors know, you can't blame the author for what the publisher puts on the cover. In this case, the front cover advertises the features of the included diskette. At the top of the feature list is "Time-saving batch files." In fact, the diskette contains only debug scripts and EXE files. I looked in vain for a single BAT file. If you want to use the batch files that Jamsa presents in the book, you'll have to type them in.

If this seems like a lot of nitpicking, the point is that if the cover and two of the first five tips in the book contain this kind of confusion, you have to be cautious using anything in the book. For example, tip 456 refers you to a program on the companion disk that lets you delete an entire directory tree. I certainly hope that there are no errors in it. Since it's in an EXE file, I couldn't read it to check it.

Many of the tips in the book simply point out useful facts or quirks. For example, tip 672 explains how to use the @ character and tells you why you don't need to start your batch files with @ECHO OFF. Tip 674 warns you of unintended consequences of using redirection or pipe operators in REM statements (comments). DOS doesn't automatically ignore them. Using IF commands to achieve conditional redirection fails similarly, as tip 718 explains.

The book groups tips into categories. System, Memory, and Keyboard account for the first 339 tips. Disk, Directory, and File account for nearly another 300. Batch and Shell make up the next 140 or so, and Hardware, Printer, and Maintenance account for almost all of the remainder.

Jamsa notes the fact that a run of 10,000 books requires the cutting of 500 trees. He pledges to donate $255 to The Basic Foundation for each 10,000 copies printed. That $255 will pay for the planting of 1,001 trees.

You can contribute to the tree-saving effort by looking this book over carefully in the bookstore; you may decide not to buy it.

Assembly language lives on

Macro Magic with Turbo Assembler, Jim Mischel (Wiley, New York, 1992, 363 pp. and a 5.25-inch diskette; $39.95)

Macros don't get enough attention. Assembly language has fallen out of fashion, so the motivation we once had to produce excellent assembly languages is gone. Microsoft's MASM and Borland's Turbo Assembler cannot compare with Digital Equipment's Macro-11, a macroassembly language available in the early 1970s for the PDP-11 minicomputer.

Given the limitations of current macroassemblers, however, Mischel has done an excellent job of showing how to access their power. Furthermore, as Jeff Duntemann of *PC Techniques Magazine* says in his foreword to the book, this is the only book ever written entirely about the subject of assembly macros for the PC. For that reason alone, even forgetting the book's excellence, you should buy this book if you think you will ever have to write or read an assembly language program for the PC.

Just for comparison, though, you might try to find a copy of *Macro Processors and Techniques for Portable Software* by P.J. Brown (Wiley, 1974). It gives a good summary of techniques people used in the good old days, when there were real macroassemblers.

Software Report

Transputers and databases

David K. Kahaner

US Office of Naval Research, Far East

kahaner@cs.titech.ac.jp

I have noticed that universities in Asia looking for inexpensive ways to move into parallel processing often develop systems from transputers. This elegant little processor is produced in a variety of versions by the UK company Inmos (newest is the T9000) with respectable computing power. (The T800 has about 2-Mflops, or 13-MIPS, peak performance.) On one chip is a CPU, a floating-point processor, built-in support for interprocessor communication, and microcoded context switching. Transputers have been configured as boards for PCs and in large parallel configurations with peak performance of several hundred Gflops. They can be physically connected by four interprocessor communication links. The communication is not exceptionally fast, but it can be performed simultaneously with calculation. The transputer's unique Occam language appears to combine parallel C with process control.

The main interest in transputers appears to be that a user can put together a system of one or more and get it up and running very easily. The four physical links make a two-dimensional mesh natural, but with software, other kinds of interconnections can be simulated, such as rings, pyramids, hypercubes. Thus users can experiment with a variety of parallel processing considerations, and several different companies produce general- and special-purpose parallel computers that are based on transputers. Not unexpectedly, transputer use is highest in the UK and the rest of Europe, but user communities exist in many other countries.

Discussions with Japanese scientists suggest that transputer activities are widely dispersed in Japan and especially focused on applications. Nevertheless, within the computer science community transputers are not at the center of attention, although this is false for those institutes that have emphasized work on them. (I would like to hear some opinions about that point.) Weaknesses center on performance degradation with scale-up to larger systems, performance a bit behind other systems at the same point in time, weak compilers other than Occam, and general complaints about difficulties using Occam. (One scientist commented that the T9000 would have been impressive if it were available a year and a half ago.)

Transputer/Occam conference. Attending this international conference in Tokyo last summer were approximately 100 scientists, perhaps 90 of whom were Japanese. Transputer meetings always seem (to me) a bit different from meetings about other parallel computers, because they attract an interesting cross section of users who are employing transputers to solve a variety of practical problems. At this meeting, the eclectic applications were obvious: data acquisition, control of power converters, VLSI logic simulation, car navigation, underwater acoustic communication, and others. Some papers discussed numerical computation including Cholesky decomposition, FFTs, a 2D particle-in-cell (PIC) for plasma simulation, a parallel Lax-Wendroff algorithm, and a parallel implementation of 0-1 knapsack problems.

For almost all of these papers, techniques already exist to solve the problems addressed, or parallel algorithms are already known. The main emphasis here was to obtain an efficient implementation. For example, the FFT paper deals with an algorithm for implementing a 1D or 2D FFT on an eight-neighbor processor array. Such an array is obtained in software using the four communication links on each transputer. (The discrete Fourier transform is developed in pow-

0272-1732/92/1200-0088$03.00 © 1992 IEEE

ers of four, rather than two.) The linear algebra paper describes a variety of experiments on variously banded systems. Few papers focused on computer science, message routing, and reduction by message passing. Two very interesting papers related to constraint satisfaction (using continuous and fuzzy variables) and multiagent planning. In these cases, the transputer is not a key ingredient, and Occam is simply used as an implementation language.

What makes transputers practical and intriguing for most of the speakers was that real parallel computing could be done with very small systems, typically only a few transputers. For example, the PIC paper deals with a problem that routinely is tasked to the largest supercomputers, involving a large Poisson solver and many particles. Eight transputers (maximum performance possible from the hardware was about 15 Mflops) and a total of 1,000 particles were used. The authors were disappointed with the parallelization performance because the algorithm required too much waiting time between computation. A more positive result was obtained from a parallel implementation of the modified (explicit) Lax-Wendroff method on a 2D grid. The application is to models of ionized gas in the workstations, but US companies such as Sun Microsystems are also well represented.

Multimedia databases. In terms of infrastructure technology, the most significant advances in Japanese efforts were in multimedia technology. Unique advances are being made by building database systems that exploit the Japanese strength in electronic devices. These include high-capacity optical and magneto-optical storage, document and image scanners, and image presentation on standard video, high-definition video, and computer-driven monochrome and color fax equipment. These capabilities are not yet well integrated into networks, and standards are lagging. The Japanese ISDN network is poised for a major expansion in bandwidth (from 256 Kbytes to 4 Gbytes). The availability of such links will further motivate this direction and cause pressure for integration of advanced multimedia, especially image technology, into databases.

Today Japan depends significantly on foreign database management system technology. Most of the vendors of these database management systems plan to provide support for the management of large, variable-size data elements, as needed for multimedia database management. It will depend on the effectiveness of these extensions whether established DBMSs will be used for the multimedia services of the future. Otherwise, developers of multimedia systems will need to develop their own DBMSs. The availability of standards such as SQL and Ada makes entry of new DBMSs that satisfy these standards feasible. Even if they are less mature, having multimedia capability can be a decisive factor in the market.

Intermediary solutions do exist. Conventional DBMSs can reference images in distinct files for images and large objects, and these can be accessed indirectly. However, such solutions are more complex to manage and are likely to be intermediate solutions. Furthermore, if pattern matching or associative access to image and voice data becomes a reality, the indirect approach will no longer be feasible.

It is becoming understood that eventually access to multimedia databases will be required. Associative access means finding an image that "looks like this image" or that contains features "like these." Speech files can be interpreted for voice print identification as well as contents. Research into this problem is in the early stages, both in the US and in Japan, so its relative success cannot now be assessed. Japanese efforts have focused on neural-net technology, which is likely to be quite effective for the simpler matching problems but may not deal well with feature-based searching. The availability of excellent technology in Japanese laboratories reduces their entry cost for researchers interested in this field. If this research direction either catches the interest of Japanese industrial research, or if academic research in this field finds support, rapid progress is possible.

[David Kahaner is on assignment with the US Office of Naval Research, Far East. His comments are his own; they do not express any official policy.]

New Products

Send announcements of new microcomputer and microprocessor products to Managing Editor, IEEE Micro, *PO Box 3014, Los Alamitos, CA 90720-1264.*

Joe Hootman

University of North Dakota

DSP components

Converter serves seismic/geophysical market

Used together, the CS5322 and the CS5323 form a 24-bit, variable bandwidth, delta-sigma A/D converter chip set for seismic, geophysical, and low-frequency passive sonar applications. The 28-pin PLCC set uses 88-mW power and offers instantaneous dynamic range (120 dB at 411 Hz) and low signal-to-distortion performance (110 dB at 411 Hz). These features allow more information to become available for computer analysis and significantly reduce the time for field calibration per channel while allowing the use of smaller geophone arrays. A monolithic digital FIR filter with programmable decimation, the CS5322 provides antialiasing for the CS5323 modulator output. The CS5323 monolithic CMOS A/D converter measures signals between direct current and 1,500 Hz. *Crystal Semiconductor; from $269.70 (100s).*

Reader Service No. 10

A complete system with flexibility

The Versatile Array Signal Processor-1000 provides the hardware and software elements necessary for development and implementation of 2-Gflops multiprocessing applications. VASP is suitable for applications that require more than board-level VME devices can provide. A flexible design permits the hardware to be configured to match the algorithm, and the three special-purpose board-level building blocks (the GSP, IOP, and TPM boards) are highly scalable. VASP is a standard 6U VME form factor, 19-inch rack-mount system with a proprietary +100 Mbytes/s system bus to connect processing elements and avoid a VMEbus communication bottleneck. *Spectrum Signal Processing.*

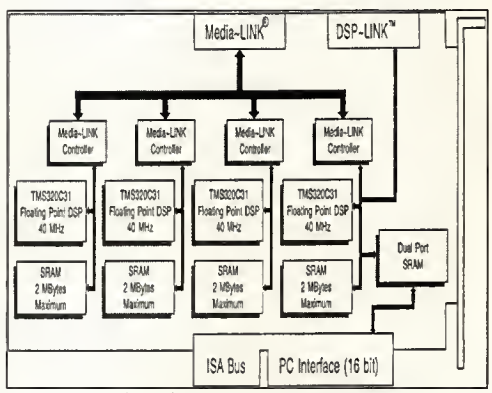
Reader Service No. 11

Boards support array, image processing

Two Media-Link boards support image processing. The MLQC31 array-processing plug-in board provides 160-Mflops peak throughput and supports desktop applications such as simulation, modeling, image processing, and radar. Based on the TMS320C31 floating-point processor, the QC31 features four independent processing elements, each with a Media-Link interface to the other three. The DSP-Link and ISA bus access augment one of the elements. The board can stand alone in a PC AT or be combined with other Media-Link products.

The second Media-Link board accepts, stores, processes, and outputs video images. MLVB contains a TMS34020 graphics processor, a 40-MHz TMS320C31 DSP, video A/D and D/A converters, and a Media-Link controller on one PC-AT card. This card can capture medium-resolution images in digital form and process them with the on-board DSP. *Spectrum Signal Processing.*

Reader Service No. 12



Spectrum Signal Processing's MLQC31

Sample at 40 MHz

A 24-bit, fixed-point chip set with support materials speeds real-time signal processing

0272-1732/92/1200-0090$03.00 © 1992 IEEE

tasks as well as system design. The LH9124 DSP includes four 24-bit complex data paths, six 24-bit multipliers, and dual 60-bit accumulators. The LH9320 address generator has more than 150 embedded sequences. The set integrates key building blocks with built-in, high-level DSP functions. A multiport architecture eliminates external multiplexing and speeds data throughput, implementing complex radix-16 butterflies in 400 ns and a 1K complex FFT in 80 µs. Cascading three LH9124s achieves a 40-MHz sample rate. A typical system consists of one DSP and three address generators, each supporting corresponding memories. *Sharp Electronics; $1,200 (LH9124-40), $250 (LH9320-40) (100s).*

Reader Service No. 13

Two functions, one card

The WAAG III board acquires data and generates arbitrary waveforms on the same IBM-compatible card. Features include either 25-MHz dual-channel or 50-MHz single-channel acquisition, 64K expandable memory, a two-range input attenuator, simultaneous sampling on both channels, and a segmented memory mode that allows repeated burst-mode acquisition. As an arbitrary waveform generator, the WAAG III provides one-channel output. *Markenrich; $1,495.*

Reader Service No. 14

Digitize signals at 120 Msamples/s

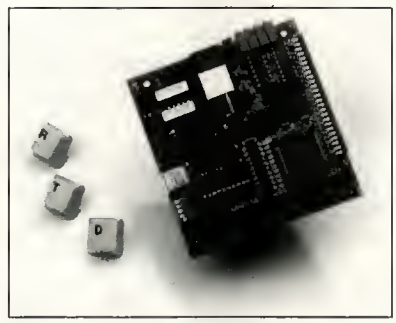
The 8-bit TDA8716 A/D converter digitizes signals at a rate of 120 Msamples/s, making it suitable for demanding professional applications, while using 780 mW of power. Featuring an input capacitance of less than 13 pF, the 32-lead surface-mount device or conventional dual in-line converter interfaces to the sample-and-hold circuits required to capture very high bandwidth signals. *Philips Semiconductors; $100.*

Reader Service No. 15

A 14-bit ADC for the PC/104 bus

The 14-bit DM412 A/D converter board for the PC-104 bus supports embedded applications. It includes eight input channels and software-programmable gains of 1, 10, 100, and 1,000. Based on the Analog Devices AD679, the 3.6 × 3.8-inch converter contains a closely matched, integral sample-and-hold circuit to ensure accurate digitization of dynamic signals to 14-bit resolution in 8 µs. The board comes with a diagnostics disk containing sample programs in Turbo C and Turbo Pascal. *Real Time Devices; $589.*

Reader Service No. 16



Real Time Devices' DM412

PC board accepts eight channels

Combining low distortion, phase coherence, and real-time error prevention, the DT2833 simultaneous sampling board preserves signal integrity and increases data accuracy. Designed for the PC AT and compatibles, the DT2833 samples up to eight differential analog input channels simultaneously, to a maximum sample throughput rate of 250,000, with 12-bit resolution. For DOS users, a device driver and tool kit package is included with each board; for Windows users, the Global Lab Data Acquisition Library, a Windows 3.0 Dynamic Link Library, is free if ordered with the board. *Data Translation; $2,595.*

Reader Service No. 17

Quad converter includes bus interface

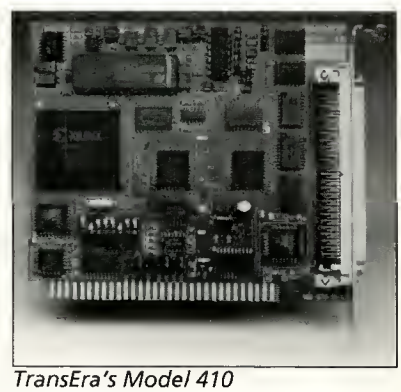
A quad 12-bit D/A converter with bus interface options, the DAC4813 supports industrial data I/O, test instrumentation, ATE, and process control applications. Its voltage output amplifier is capable of swinging ±10V while operating power-supply voltages of ±12V to ±15V. The bus interface features a 12-bit port with an input buffer latch and a holding latch for each D/A converter. A reset function allows the user to reset D/A inputs to bipolar zero. DAC4813 comes in a 0.6-inch, 28-pin plastic dual in-line package. *Burr-Brown; from $29.95 (100s).*

Reader Service No. 18

Transfer data continuously

The Model 410 data acquisition plug-in board provides software drivers for use with HTBasic and Microsoft or Borland C. Features include 16 single-ended or eight differential analog input channels, an input range of ±5 volts, and a successive approximation unit A/D converter with 13-bit resolution and 50,000-sample throughput rate. The architecture enables pseudosimultaneous sample-and-hold and triggering, and setting the number of pre- and post-trigger data points in each scan. With a 4 × 5-inch card size and an edge connector for an 8-bit bus, the surface-mount board operates in XTs through 486 PCs and compatibles. *TransEra; $495.*

Reader Service No. 19



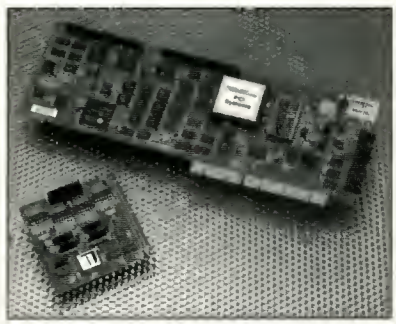
TransEra's Model 410

Expand EISA boards

Users of PCI-20501C-1 EISA data acquisition boards can add 32 single-

ended analog input channels using two PCI-20368M-1 modules. With a 1-µs advance rate, this plug-in module supports sampling rates up to 1 MHz. Daisy-chaining seven modules makes 128 input channels available. Features include high-speed buffer amplifiers for each channel, the option of software-controlled channel selection or automatic channel sequencing, and the ability to independently set gains of 1 to 100 for each channel with user-installed, plug-in resistors. *Intelligent Instrumentation; $550.*

Reader Service No. 20



Intelligent Instrumentation's PCI-20368M-1

Zero-drift op amp promises low noise

The LTC1250 chopper-stabilized, zero-drift operation amplifier reduces typical 0.1-Hz to 10-Hz noise to as low as 0.65 microvolts peak-to-peak, while providing an output swing of 4.2V into a 1K-ohm load and operating on a single 5V supply up to ±8V. With sample-and-hold capacitors included on board, the LTC1250 recovers from overload in 1.5 ms. This op amp is best used with low-impedance sources, especially bridge transducers. The part comes in military and commercial versions in 8-lead plastic or ceramic dual in-line packages and 8-lead surface-mount packages. *Linear Technology; from $2.90 (100s).*

Reader Service No. 21

Large memory plus versatility

The Cyclops DSP-C40 board for 80286, 80386, and 80486 PCs combines a large memory with versatile I/O and multiprocessing capabilities. Based on the 32-bit, floating-point TMS320C40, Cyclops occupies a single PC-AT slot and offers a peak integer performance of 275 MOPS and a peak floating-point performance of 40 Mflops. The DSP board provides 32 Kbytes of dual-ported memory (memory or I/O mapped) that is shared with the PC, and includes up to 64 Mbytes of DRAM and 6 Mbytes of SRAM (zero or one wait state). To support multiprocessor configurations, Cyclops makes the TMS320C40's six 20-Mbyte/s parallel ports available via six connectors. DT-Connect, a 16-bit, 20-Mbyte/s parallel interface, gives designers access to high-speed video devices. Software support includes TI's ANSI-compliant C compiler. *Ariel; $5,995.*

Reader Service No. 22

Video/computer components

PC multimedia tool announced

PC users can let TelevEyes convert from VGA to recordable composite video. TelevEyes uses an external module that connects between the computer's output and the monitor, outputting an NTSC composite video signal of whatever is on a VGA screen. A composite video jack then connects to a VCR, projector, light panel, or other display device. The TSR control software, which supports computer text and graphics display modes up to 640 × 480, features simultaneous composite and computer display, accurate NTSC color mapping, and flicker-free composite video output. *Digital Vision; $299.95.*

Reader Service No. 23

Converts to broadcast quality

The Super Encoder board lets users convert their PC's VGA output to broadcast-quality video. When operated in conjunction with the Super VideoWindows ISA board, the converter allows users to create and modify full-motion, video-based presentations on the PC and then record them onto videotape. Using digital encoder technology from TRW, the board creates output in an NTSC, PAL, or S-VHS signal. Super Encoder provides programmable hue settings and on-board lookup tables for VGA palette mapping, gamma correction, and other pixel processing. A direct digital mode accepts digital VGA from a VGA card's feature connector. In 24-bit RGB mode, an analog VGA signal is input to the Super Encoder, digitized, and input to the board's digital encoding section. The board requires MS-DOS or PC-DOS Version 3.3 or higher. *New Media Graphics; $595.*

Reader Service No. 24

RGB/Videolink adds RS-232 port

The RGB/Videolink 1600U video scan converter, which transforms high-resolution computer graphics to television format, now includes an RS-232 port to control all functions directly from a computer. The 1600U automatically synchronizes all computer displays with 20- to 90-kHz horizontal scan rates, including those from desktop computers and workstations. The converter, which accepts both interlaced and noninterlaced inputs, measures the input signal's frequencies and sets appropriate parameters. The 1600U's direct interface to display equipment accepts signals up to 32 kHz. Offering a double-rate 31.5-kHz output simultaneously with the broadcast video, the 1600U can map any number of input lines to any number of output lines. Features include a zoom function, antialiasing, 24-bit color processing, real-time operation, third-generation DSP circuitry, three levels of filtering, and a built-in linear keyer. *RGB Spectrum; $19,495.*

Reader Service No. 25

Mac video conferencing

Desktop Visual Communications products for Macintoshes provide both real-time and store-and-forward communications of voice, video, data, and documents over ordinary telephone

continued on p.94

Special products

Smart computer speeds at 128 Mcps

The RN-200-based neurocomputer learns by example at 1.5G connection updates/s and operates without complicated software, allowing speeds of 128M connections/s. A CMOS channel-less gate array with 200,000 gates allows one LSI RN-200 chip to fabricate 256 synapses, compared to eight previously, with the use of a 0.8-micron design rule (circuit minimum line width). The company offers a prototype to systems development engineers. *Ricoh.*

Reader Service No. 26

Tabletop system uses 3.5-inch drives

A redundant array of inexpensive drives (RAID) that incorporates 3.5-inch disk drives provides up to 15 Gbytes of data capacity. The RAIDstor-T3 is a multiplatform data storage system for software-transparent operation on the SCSI port of computers from Macintosh, Digital Equipment, Hewlett-Packard, IBM (RS/6000, PC and compatibles), NeXT, and Sun Microsystems. The system's five drives, independent power supplies, fault isolation, and on-line repair capabilities make it a near-fault-tolerant unit.

The compact tabletop system supports all three levels of RAID implementation, and its SCSI interface transfers data at 20 Mbytes/s. The RAIDstor-T3 employs the same superscalar, 25-MHz Intel 80960-CA microprocessor as the larger RAIDstor systems and is available in a driveless package. *Unbound; from $23,800 to $65,700.*

Reader Service No. 27

Neural network tool supports Iris family

NeuralWorks Professional II/Plus is now available for use on Silicon Graphics' Iris family of RISC-based systems. This neural network development tool also supports the Macintosh, PC, and compatibles, plus RS-6000, Silicon Graphics, Hewlett-Packard, Sun, and DEC workstations. The tool provides prototyping and concept testing of neural network designs for a variety of data-intensive, time-sensitive, and quality-dependent applications.

Written in C, Professional II/Plus features a graphical user interface, an open architecture, and support for 22 major neural network types. Features include ExplainNet, which tells users how a neural network came to a particular conclusion, and FlashCode, which interprets backpropagation networks created in Professional II/Plus and then generates C source code to create embedded applications. *NeuralWare; $4,995.*

Reader Service No. 28

RISCs, embedded processors aided

The Microprocessor Analysis Package (MAP), operating with the Configurable Logic Analysis System (CLAS) family, expands support for Motorola RISC and embedded microprocessors. MAP captures microprocessor activity and displays disassembly in standard mnemonics or as timing diagrams and state listings. MAP, which supports 20-ns instruction access bus rates, automatically configures the analyzer to match the target microprocessor architecture. A high-impedance probe connection ensures against disruption of the system being examined.

The CLAS analyzer, with a 19-inch display, shows all typical disassembly information in one window, and up to 13 windows may be opened simultaneously. Biomation Trace Control provides 15 levels of sequential event recognition and selective recording to identify and capture complex combinations of conditions. The CLAS analyzer may be configured from one to four independent instruments and supports simultaneous multiprocessor disassembly. *Biomation; from $20,900.*

Reader Service No. 29

Biomation's CLAS Microprocessor Analysis Package

DCS captures images for desktop

The DCS 200 digital camera captures images at 1,012 × 1,524 resolution for use in desktop computers. A special back in the Nikon 8008s camera body contains a CCD array. Features include rapid auto focus, exposure control, motorized advance, and lens flexibility. An SCSI port links the camera directly to Macintosh and PC-compatible computers, and the package includes drivers for Adobe Photoshop and Aldus PhotoStyler. Four models of the camera are available, with either color or monochrome capability and the capacity to store one or 50 images. *Kodak; from $8,495 to $9,995.*

Reader Service No. 30

lines. This technology combines collaborative document-sharing software, screen-based telephone management software, and the proprietary Vector Adaptive Transform Processing hardware and software. VATP, which provides high compression ratios using standard international modem technologies, allows low-cost implementations of motion, color video, and high-quality audio for desktop communications, according to the manufacturer. Its programmable hardware supports various video- and image- industry standards. DVC, which supports QuickTime and Apple's OCE, combines two NuBus cards and ShareVision software with a color video camera and a Norris Ear Phone. *ShareVision.*

Reader Service No. 31

Powerful set has multimedia applications

The CL-PX2070 and the CL-PX2080 process and display multiple video streams for personal-computer and video-teleconferencing applications. The CL-PX2070 programmable DVP handles multistream video. The CL-PX2080 MediaDAC chip digitally mixes and simultaneously displays graphics with multistreams of live video, allowing users to easily view and manipulate video windows.

Implemented together, the devices yield 1,024 × 768-pixel, true-color video systems. The DVP accommodates applications that require several concurrent video and graphic sources, features an advanced frame buffer controller and a programmable ALU, and offers bidirectional data paths between its ports and internal control and processing functions. Each device is packaged in a 160-pin plastic quad flat pack. Volume production is planned for early 1993. *Cirrus Logic; $85 (CL-PX2070, 1,000s),* $65 (CL-PX2080).

Reader Service No. 32

Advances in video encoding

A real-time, full-motion, color video compression/decompression system offers 640 × 480-pixel resolution performing at an SMPTE-standard 30 frames/s. According to the manufacturer, the Pro-Motion Video PC board set and Video Developers tool kit offer significant advancements over the current MPEG architecture for video encoding. The system uses the Four Square Transform compression algorithm, which offers real-time compression/decompression with zero latency, high picture quality even with high motion or a busy background, and higher subjective picture quality compared to MPEG encoding. The flexible system offers scalable video quality, customized programs, and 25 video compression parameters. The boards incorporate DMA bus mastering, accept input from S-Video or NTSC devices, and work with the PC ISA standard. The minimum configuration is a 12-MHz 80286 computer with two free 16-bit slots running DOS 4.0 or higher.

The developer tool kit provides the information and tools needed to produce sophisticated video applications and includes the 80286 assembler source code for the TSR video capture and display drivers. *AWA; $7,900 (board set), $895 (tool kit).*

Reader Service No. 33

Scientific/design software

Solutions tested without synthesis models

Three Test Design Expert products, all operating without a synthesis model, join the TDX Step to form a series of test automation stepping stones. They share database and CAE-interface compatibility.

Full Scan offers a stand-alone version of the scan-insertion, rule-checking, and dynamic vector-compaction technology available in the original product. The FPGA test solution, based on Step's technical innovations, contains different algorithms for sequential test generation that use only the gate-level netlist. Partial Scan is a fully functional sequential test generation solution with integrated design-for-test support. For users with a complex circuit and no synthesis model, the Partial Scan software develops a high-quality partial scan test that works around scan rule violations, embedded RAMs, and critical timing storage elements.

The TDX family is available on Sun Sparcstations, the HP700 series, and RS6000 computers. *ExperTest; $36,500 (TDX Full Scan), $45,500 (TDX FPGA), $65,500 (TDX Partial Scan).*

Reader Service No. 34

Animate scientific graphs on a PC

Covis displays animated views of data in real time. Users can create Nchannel, 2D and 3D snake, contour, vector, wireframe, points, and 2D and 3D programmable graphs. The PC program also runs on PS/2s or compatibles running DOS 2.0 or higher with 512 Kbytes of RAM and a hard disk. It supports Hercules, EGA, ATT 6300, MCGA, VGA, and Super VGA graphics cards and can import data from ASCII, dBase, Excel, Lotus 1-2-3, Quattro, and other data files. The manufacturer recommends an S3-based Super VGA card for optimal perfomance. CoVis can use (but does not require) a Microsoft-compatible mouse. *CoHort Software; $395.*

Reader Service No. 35

Reader Interest Survey

Indicate your interest in this department by circling the appropriate number on the Reader Service Card.

Low 189 Medium 190 High 191

Product Summary

Joe Hootman
University of North Dakota

Manufacturer	Model	Comments	R.S.#
Chips			
Advanced Micro Devices	Am28F010A/ 20A, Flash memories	Available with guaranteed 100,000 write endurance cycles (minimum) and automated program and erase operations, these 12V 1- and 2-Mbit devices support embedded Flash disks and removable memory cards. $9.90 (28F010A), $20.75 (28F020A) (100s); 32-pin PLCCs, PDIPs, and TSOPs.	80
Data I/O	Universal Programmer family	Current 3V erasable PLDs, EPROMs, and Flash memories can be programmed, verified, and tested with V_{IH} and V_{IL} levels and tolerances that match device specifications for preinstallation circuit performance analysis. The programmers use software-controlled pin drivers to provide 12-bit DAC- and ADC-controlled signals. From $3,450 to $9,995 for base units.	81
Matra MHS Electronics	80C5X controllers	Operating at 16 MHz over $2.7V < V_{CC} < 5.5V$, the two 3V microcontrollers consume 12 mA of power. The devices' static cores allow engineers to reduce power consumption further by lowering the clock rate. From $9.50 (5,000s)	82
Motorola Computer Group	MVME197 RISC	Single-board computer based on the 88110 Symmetric Superscalar RISC microprocessor supports simulation, telecommunications, compute-intensive applications, and various connectivity requirements, including FDDI, Ethernet, SCSI, and graphics. Promising over 70 SPECmarks at 50 MHz, the 197 offers floating-point and integer performance with its six ASICs, Unix System V/88 4.0 support, and VMEexec development tools. From $9,995 (sample quantities); volume shipments 1Q93.	83
Unitrode Integrated Circuits	UCC3883/85 chip set	BiCMOS PWM chip set implements ISDN-compatible power supplies operating at more than 50% efficiency with a 25-mW load. The 3883 peak-current mode controller features zero-power start-up, restricted-mode detection, and low-quiescent power for CCITT needs. The 3885 secondary-side regulation IC provides feedback control voltage and oscillator synchronization data to the controller via an isolation pulse transformer. $2.42 (UCC3883), $2.46 (UCC3885) (1,000s).	84
Telecommunications			
Paladin Software	MicroTAP 2.1 monitor/ debugger	Debugging, data capture, and analysis tool supports computer programming, manufacturing, industrial automation, and multimedia applications. The new version of DataScope is a serial-line monitor that includes context-sensitive Hypertext, Hypersetup, user-alterable multitasking window displays, and oscilloscope-like signal event tracing. Data and signal events are time-stamped to the microsecond. $299, including cable, connectors, and manual.	85

FOR DISPLAY ADVERTISING INFORMATION, CONTACT:

Southern California and Mountain States: Richard C. Faust, Douglas C. Faust, 24050 Madison Street, Suite 101, Torrance, California 90505; Tel: (310) 373-9604; Fax: (310) 373-8760.

Northern California and Pacific NW: William W. Hague, 9017 Peacock Hill Road, Gig Harbor, Washington 98332: Tel: (206) 858-7575; Fax (206) 858-7576.

East Coast: Gail A. Frank, Nancy Inserra, 82 Bethany Road, Suite 4, Hazlet, New Jersey 07730; Tel: (908) 264-1100; Fax: (908) 264-4340.

New England: Paul Gillespie, PO Box 6444, Holliston, Massachusetts 01746; Tel: (508) 429-8907; Fax (508) 429-3285.

Advertising Manager: Heidi Rex, 10662 Los Vaqueros Circle, Los Alamitos, California 90720-1264; Tel: (714) 821-8380; Fax: (714) 821-4010.

For production information, conference, and classified advertising, contact Marian Tibayan.

IEEE MICRO, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, California 90720-1264; Tel: (714) 821-8380; Fax: (714) 821-4010.

IEEE COMPUTER SOCIETY®

A member society of the
Institute of Electrical and Electronics Engineers, Inc.

THE FOLLOWING INFORMATION IS AVAILABLE:

Contact the Publications Office;
to facilitate handling, please request by number.

- Membership application, student #203, others #202
- Periodicals subscription form for individuals #200
- Periodicals subscription form for organizations #199
- Publications catalog #201
- Compmail electronic mail brochure #194
- Technical committee list/application #197
- Chapters lists, start-up procedures #193
- Student scholarship information #192
- Volunteer leaders/staff directory #196
- IEEE senior member grade application #204
  (requires ten years practice and significant performance in five of those ten)

To check membership status or report a change of address, call the IEEE toll-free number, 1-800-678-4333. Direct all other Computer Society-related questions to the Publications Office.

PURPOSE

The IEEE Computer Society advances the theory and practice of computer science and engineering, promotes the exchange of technical information among 100,000 members worldwide, and provides a wide range of services to members and nonmembers.

MEMBERSHIP

Members receive the acclaimed monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

PUBLICATIONS AND ACTIVITIES

Computer. An authoritative, easy-to-read magazine containing tutorial and in-depth articles on topics across the computer field, plus news, conferences, calendar, interviews, and product reviews.

Periodicals. The society publishes seven magazines and five research transactions. Refer to membership application or request information as noted at left.

Conference Proceedings, Tutorial Texts, Standards Documents. The Computer Society Press publishes more than 100 titles every year.

Standards Working Groups. More than 100 of these groups produce IEEE standards used throughout the industrial world.

Technical Committees. More than 30 TCs publish newsletters, provide interaction with peers in specialty areas, and directly influence standards, conferences, and education.

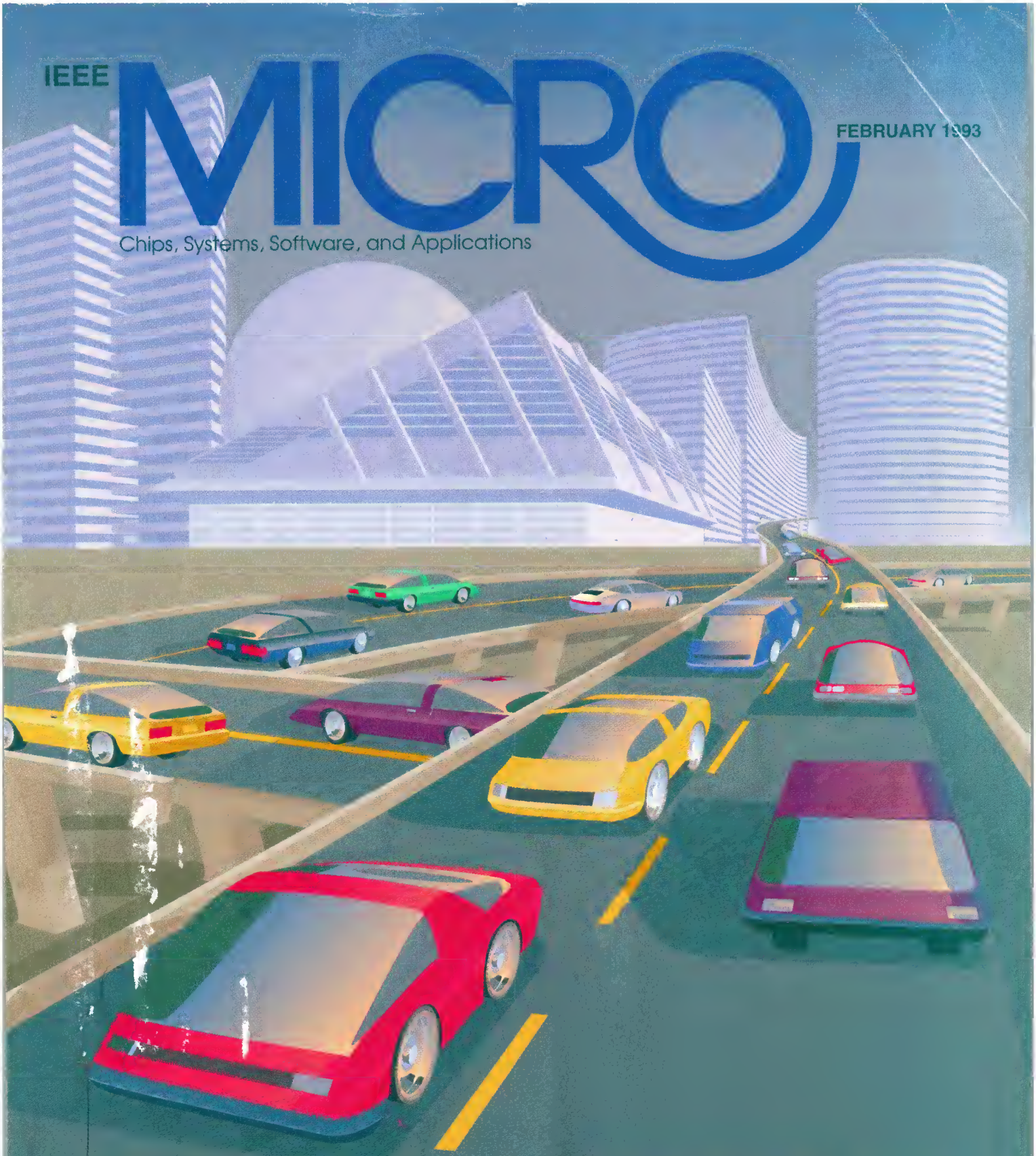
Conferences/Education. The society holds about 100 conferences each year and sponsors many educational activities, including computing science accreditation.

Chapters. Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

OMBUDSMAN

Members experiencing problems — magazine delivery, membership status, or unresolved complaints — may write to the ombudsman at the Publications Office.

IEEE
MICRO
Chips, Systems, Software, and Applications
FEBRUARY 1993



AUTOMOTIVE ELECTRONICS

Traffic Management • Intelligent Control
Roadside Information • Vision Enhancement • Safety and Reliability

IEEE COMPUTER SOCIETY
THE INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS, INC.

IEEE MICRO

Published by the IEEE Computer Society

Volume 13 Number 1 February 1993

FEATURES

THEME ARTICLES

SPECIAL REPORT

Cover: Alexander Torres

Circulation: *IEEE Micro* (ISSN 0272-1732) is published bimonthly by the IEEE Computer Society, PO Box 3014, Los Alamitos, CA 90720-1264; IEEE Computer Society Headquarters, 1730 Massachusetts Ave., NW, Washington, DC 20036-1903; IEEE Headquarters, 345 East 47th St., New York, NY 10017. Annual subscription: $23 in addition to IEEE Computer Society or any other IEEE society member dues; $42 for members of other technical organizations. This journal is also available in microfiche form.
Postmaster: Send address changes and undelivered copies to *IEEE Micro*, PO Box 3014, Los Alamitos, CA 90720-1264. Second-class postage is paid at New York, NY, and at additional mailing offices. Canadian GST#125634188.

DEPARTMENTS

Call for Articles, p.10, 56; Reader Service cards, p. 80A/B; Advertiser/Product Index/ Moving Coupon, p. 88; Computer Society information, cover 3

IEEE Computer Society
PO Box 3014, Los Alamitos, CA 90720-1264
(714) 821-8380

* Submit six copies of all articles and special-issue proposals to Dante Del Corso, Dipartimento di Elettronica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129 Torino, Italy; phone +39-11-556-4044; Compmail: d.delcorso; Internet: delcorso@polito.it; Bitnet: delcorso@itopoli

or

** Maurice Yunik, Dept. of Electrical Engineering, University of Manitoba, Winnipeg. Manitoba R3T 2N2, Canada; phone (204) 474-8517; yunik@eeserv.ee.umanitoba.ca

Author guidelines for article submission are available from the West Coast office (address at left).

From the Editor in Chief

Check point two

IT IS TIME to check the balance sheets again: One more year of work has been accomplished, and we still have work to do and plans to make for the future.

For the electronic industry (as for many others), 1992 could not be classified as a "booming" year. While technology continued to produce more and more impressive results, companies continued to lay off people. This was a common denominator in the US and Europe last year, but now some hope of a trend reversal is appearing, at least in some areas.

These economic problems also impacted the life of technical publications, like our magazine. People have less time to write (unpaid) articles and to review those written by other authors. Subscriptions do not increase. (This last statement is formally correct, even if someone could say I express the situation from an optimistic point of view.) The Editorial Board and the editorial staff worked hard to keep *Micro* going through 1992, maintaining or improving the level of service provided to readers. You can judge whether we succeeded.

In particular, the efforts of managing editor Marie English and the use of new technologies at the Los Alamitos office allowed us to keep publication costs low. Low costs are critical in delivering the number of pages you are accustomed to—and even more to appreciate, since for much of 1992 the *Micro* editing staff went from two people on one magazine to one person on two magazines.

The main scope of *Micro* is to bring useful information to readers, but in our field the value of information decreases with time. Therefore, the efforts of the Editorial Board in reducing the review time of manuscripts submitted for publication continued in 1992. I am proud to say that the average delay from submission to acceptance (or rejection) is now around three months. Those of you who are familiar with technical publications can appreciate the value of this figure. Referees play a key part in the review process. Authors know how valuable are comments and suggestions from other experienced people. To acknowledge this work, each year *Micro* will publish the list of referees who contributed to the previous year's issues. We heartily thank each of the 1992 referees you will see listed here; they take time to see that *Micro* continues to be the well-received magazine that it is.

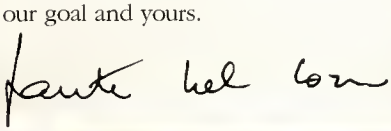
What plans do we have for the coming year? We plan to keep and increase our efforts at disseminating information that is useful to microsystems designers. We plan to place more emphasis on education (we are looking for good tutorial articles) and on standards. This last theme is a warhorse for *Micro* and is becoming more and more important as all markets become worldwide. Steve Diamond, the new editor for the Micro Standards department, will address the technical aspects and motivations, both of established standards and of the many efforts under way.

The application areas of microelectronics and microsystems are continuously expanding, and *Micro* plans its content to cope with this process. In 1992 you could read special theme issues on the latest microprocessors (Hot Chips issue), associative memories, a snapshot of the European microprocessor industry, video chips,

0272-1732/93/0200-002$03.00 © 1993 IEEE

and special signal processors. In 1993, besides this current issue on automotive electronics, you will be able to read about packaging and interconnections, plus the latest news from the Hot Chips conference, Far East industry, and standards. Plans for 1994 are under way. Be ready for hot themes like fault-tolerant systems, optical computing, intelligent sensors, and more.

This is what we can provide to readers, but we must also receive from them. What I feel is missing is more feedback. In electronics it is well known that positive feedback (in this case confirmation that what we are doing is correct) must be kept to a minimum to avoid instability. On the other hand, negative feedback (what you do not like, what we should change) is extremely important. Please continue to make proposals and suggestions on what to add, change, or cut; making *Micro* better and better is our goal and yours.



Mailbag

(LK: liked; DLK: disliked; LTS: like to see)

October 1991

LTS: More detailed information about ICs and microprocessors.—V.D., Moscow

February 1992

LK: Am29000; LTS: DSP processors—M.E.M., Teheran, Iran [The December issue should fulfill your request.—D.D.C.]

LK: Neural network classifier; LTS: everything is OK; you are on the right path.—R.V.S., Ljubljana, Slovenia [Thanks; any suggestions for doing better?—D.D.C.]

April 1992

LTS: DEC Alpha; monograph on RISC.—P.P., Civitanova, Italy [RISC architectures are extensively covered in the Hot Chips special issues (February and June 1990, June 1991, and April 1992); DEC Alpha is coming.—D.D.C.]

LK: Micro Law (Nintendo v. Galoob)—D.S., Ottawa, Canada

LK: The R4000 and 88110 RISC reviews; DLK: not knowing what SPEC packages are used to benchmark these processors; LTS: these packages explained and Mflops rates in next review (also Linpack).—A.H., V.N., de Gaia, Portugal

LK: Motorola 88110 review and Mips R4000 processor.—S.D.K., Bandung, Indonesia

LK: Articles on RISCs; LTS: DEC's Alpha and NVAX RISC processors; DEC's Open Advantage and Open VMS.—J.F., Ljubljana, Slovenia

LK: MDP, R4000.—H.W., Bandung, Indonesia

1992 referee list

J. Acquah
D. Al-Khalili
V.K. Arora
M. Atkins
A. Bakkers
J. Biemond
J. Boillat
F. Cathoor
P. Chow
P. Civera
D.W. Clark
L. Criminiera
K. Das
M. Dean
Delogne
N. Demassieux
P. Duzy
S. Efstratiadis
M. Farrens
K. Ferens
J. Fogler
D. Frank
A.S. Gajowski
L. Galbraith
D.J. Gelet
W. Gerling
R. Goubran
F. Gregoretti
E. Grochowski
K.E. Grosspietsch
K. Gruger
A. Gupta
O. Grüter
W.S. Heath
J.L. Hennessy
H. Herbst
H. Herrmann
M.D. Hill
B. Hoefflinger
V.K.L. Huang
M.R. Ito
K. Jainandunsing
M. Johnson
D. Kakoske
W. Kinsner
P. Klarer
U. Klein
A. E. Knowles
M. Koeniger
J. Kowalczuk
Y. Ktayama
L. Litzler
Longchamp
F. Maddaleno
T. Maguire
J. McCoy
P. Michel
D.R. Miller
J. Miner
N. Miner
T. Muntean
P. Mussi
M.A. Narasimhan
A. Neufeld
M. Newman
D. Nicole
Nobuyuki
W. Pedryez
G. Piccinini
M. Pomper
P. Prinetto
D. Quammen
B. Rahardjo
T. Ramstad
R. Reetz
G. Reichart
B. Reiner
D. Renshaw
L. Reyneri
M. Ruo Roch
G. Rosendahl
W. Rosenstiel
U. Rueckert
W. Ruecker
C. Sansoé
G. Scarth
U. Schaefer
L. Spaanenburg
S. Sriram
R.E. Seviora
J.E. Smith
G. Sohi
R. Tatge
V. Thomas
T. Todesco
D. Tolsch
M. Trehel
G. Wagner
K. Waldschmidt
W.-D. Weber
P. Weissglas
N. Yagi
M. Zamboni
H.-G. Zimmer

ICs per vehicle increasing rapidly

Ware Myers

Contributing Editor

The average number of integrated circuits per automotive vehicle is now 89, up from 70 only two years ago. These numbers characterize the growth rate of automotive electronic systems, Jerry Rivard, former chief engineer, electrical and electronics, for Ford Motor Company, said in an *IEEE Micro* interview. Rivard started out in automotive electronics in the mid-1960s in Bendix Corporation's Advanced Automotive Concepts Program and later became group director of engineering for the Electronic Fuel Injection Division.

"We had cars running then with headway control, antilock braking systems, and electronic fuel injection; but we were too early," he said. "It didn't begin to happen until the mid-1970s."

Rivard headed the team that put the first electronic fuel injection system on the Cadillac Seville in 1975. In 1976 Ford asked him to organize its electronic program, and for 10 years he was chief engineer. In 1986 he returned to Bendix (Allied Signal Inc.) as vice president and group executive of Bendix Electronics. Recently he has been a consultant in the field. He is a fellow of the IEEE and the Society of Automotive Engineers and a member of the National Academy of Engineering.

Is there a difference in automotive electronic systems put on high-end cars and low-end vehicles?

You generally find functional systems, such as engine and transmission control, antilock braking, and air bag, going across all cars. Manufacturers need these systems to meet regulatory requirements like emission reduction, fuel economy, or safety. Merely giving the driver some convenience, like antitheft, electrochromic rearview mirror, keyless entry, or an exotic entertainment unit, adds costs the average driver is not willing to pay.

What is an electrochromic mirror?

An electrochromic phenomenon on the mirror darkens the reflected light responding to sensed headlights of a car approaching from the rear. Then, instead of having to reach up to the mirror and snap a button, it dims the mirror automatically.

What are the major electronic systems now found on cars?

The engine control module controls the power train. The latest version controls both the engine and the transmission. Some cars have a module for electronic-hydraulic steering that changes the gain on the steering system. An antilock braking module is catching on quickly. Other examples include a diagnostic module for the air bag system and a central module for instrumentation.

What do you see coming in the next two or three years?

The biggest growth area is the antilock braking system and extensions of it, such as traction control. You might get one wheel stuck in snow, ice, mud, or sand, where it just spins, and the wheels with traction don't move at all. Traction control transfers the torque from the spinning wheel to the wheels with traction, enabling the car to move out.

Air bags are coming on very quickly. The safety value has been proven. They will be going across all vehicles by the mid-1990s, from high to low, both driver- and passenger-side.

What is coming after that?

There is a lot of work in the industry laborato-

0272-1732/93/0200-0004$03.00 © 1993 IEEE

ries on headway warning or control. These systems emit a radar or light beam (lidar) ahead of the vehicle to sense an object you want to avoid. The early systems probably will just give the driver an audible alarm. A still more advanced system is radar speed control. It warns you of too rapid a closure rate with the car in front or of a car cutting in front of you. If you don't take action, it will close the throttle and start some braking effect.

A project at the University of California, Berkeley, is developing a system of this type with the objective of moving more traffic. On a stretch of freeway in San Diego they are running a string of 10 to 12 cars spaced about three meters apart. The lead car sets the pace. If there is any change, the system provides braking and steering functions on the following cars. The system would allow more throughput, safely on a crowded freeway.

One of the biggest problems, honestly, is not technical; it is the fear of liability. The manufacturer who creates a new technology worries about the risk of some unknown failure. With any complex system, you are going to have at least a few failures. Before a system goes out to the public in the automobile industry, it must be highly reliable.

That leads us into questions of design. How did you go about introducing new technology?

The first thing you have to understand about the automotive industry is that it has its own way of doing things. Automotive management is basically skeptical about new technology. Moreover, most of the engineers are mechanical and don't understand electronics.

I learned that you don't come in with ideas that are not well thought out. An idea on paper doesn't sell. You have to come in with something demonstrable. You have to reduce the new idea to practical practice. When you are putting a million cars on the road, you can't afford something that doesn't work well.

Semiconductors weren't very reliable in the early days. I remember the Japanese made a big impression a few years later with more reliable chips.

Yes, I had problems on the automotive side, and I had problems on the semiconductor side. At first the semiconductor people had no feel for the automotive business. When the first ICs came out, the markets in computers and the like were huge. From the time the semiconductor companies started an idea to the time it made profits was as short as a year. But the automotive industry is very slow moving. It takes five years from the time you accept an idea until you see it in practice.

Semiconductor executives who later became good friends like Bob Noyce and Gordon Moore of Intel saw the prospect of investing money with only a long-term payback. At the time they were making money on product ideas with a fast turnaround. I had to sell both sides. It took about 15 years.

Where is the industry today?

It has become pervasive. Semiconductor sales to the automotive industry are around $2.5 billion, forecast to go to $5 billion by the middle of the decade. Market analysts expect electronic system sales, now about $8 billion, to reach $24 billion by the year 2000.

To get that kind of growth, you had to do something about reliability.

In the early 1970s drivability was atrocious. Emission regulations were just coming in, and the engineers were trying to cope with them using conventional mechanical technology. As a result they were not getting performance. For instance, you might have to start and restart your car three times before you got out of your garage. Now, with the electronic systems, you don't even think about things like that. You turn the key and bang! The car starts and stays started.

You have to give credit to the Japanese. They saw the need for reliability, and they understood the fundamentals of it. In those days when we multiplied the component reliability numbers together, we ended up with system assembly figures that no one wanted to put in the car. Well, we eventually solved that problem with a lot of demanding, pushing, and cooperation. At the same time we had to control costs. The automobile industry is extremely cost-sensitive.

As you put more microprocessors in cars, you must have been putting in more software, too. What did you do about software reliability?

In 1978-80 we put our first digital ICs in vehicles. Up to that time we had only analog systems. We put software in the IC processors and ended up with huge software problems. Of course, the ICs of that period were unreliable, too. In test it was hard to tell if a problem was caused by hardware or software. Our ability to verify software in the actual installation was not good.

We gradually developed tools and methods that allowed us to check software. Today we don't see a lot of software problems.

How did you get these 5-volt electronic circuits to operate reliably in the car's noisy electrical environment?

Well, we had trouble with electromagnetic interference. We finally had to write a textbook on how to design ICs into this harsh environment, and how to interface the ICs to sensors.

Sensors, even today, are one of our biggest problems. They are the Achilles' heel of an electronic system. In the 1970s we had to use what was available from aerospace, but they had few cost constraints. We had to adapt the technology to our industry.

Sensors are still not as reliable as they should be. Most of them are overpriced by a factor of two. We don't have the accuracy levels that we need for the next generation of control systems.

Do your systems operate independently, or are they bused together?

Communications between processors are changing dramatically. It will help to look at how this field evolved. We have gone through three phases and are in the fourth phase now. The first phase was just putting electronic components like clocks and radios in the car—no connections between them. In the second phase we put in electronic subsystems that merely emulated the mechanical system that already existed. But the new system was not optimized; it didn't take advantage of the potential of electronic systems.

In the third phase we recognized that we were proliferating subsystems, getting endless complexity. We began to ask: How do you interface these systems? How do you share sensors and databases? How do you optimize them? How do you diagnose malfunctions? We were in the system-engineering phase.

For instance, wiring harnesses were getting out of control. If a car door had all the available controls, you could have 50 or 60 wires running into it—a bundle as big as your wrist. Difficult to build, package, and install reliably. Also costly. We had to move toward multiplexing.

If everybody multiplexed in their own way, we would end up with protocols that would be costly and difficult to service. So the Society of Automotive Engineers and the International Standards Organization formed committees—with Japanese participation—to standardize multiplexing.

You are not going to see the whole car multiplexed overnight. It is coming in only where needed to reduce the number of wires and connectors, to move data from one system to others that use it, or to share sensors. Multiplexing also improves your diagnostic capability. You can interrogate different systems from a central point, decide what is wrong, and show how to repair it. A sensor on the transmission, for example, tells you how fast the drive shaft is turning. You need that information for engine control and antilock braking.

That transfers the complexity back into software.

Well, there is a benefit to putting as much as you can into the software. It gives you flexibility in handling year-to-year model changes as you come to understand system needs better. It reduces the cost of making changes.

You sound as if you had confidence in the industry's ability to write error-free software.

Well, we have come a long way. One number I remember: The air bag system is 99.99999 percent reliable. That is the design value. Engine control, of course, is a lot more complex. The possibility of software errors or hardware failures is greater because the number of components is much larger.

Engine control is like running a little chemical plant.

Exactly. Not only that, but the speed of response is critical. We are up to 18 MHz on the engine control units, and the designers want higher frequency to give them better accuracy.

You mentioned a fourth phase. What is it?

It is where we look not only at the systems on the car but also at the larger system, that is, the road system or the infrastructure, that the car operates in. It is the phase the Intelligent Vehicle-Highway Systems researchers are studying.

Guest Editor's Introduction

An Electronic Copilot in Your Car?

Bernd Hoefflinger

Institute for Microelectronics Stuttgart

Electronics in the car continues to be a much debated issue. Fascination about its potential on the one hand and concerns about its invisible inner workings on the other hand, together with the impact on individual safety and freedom of action, provide a challenge probably unparalleled in any other field of applied microelectronics. With over 500 million cars on the world's roads, we certainly stretch our imagination if we think that all these units one day may have airplanelike cockpits in them. Moreover, with this or just because of this comparison, everyone of us can instantly quote many good reasons why electronic road traffic will be much more complex than electronic flying.

Appropriately, in the area of cooperative civil technology research and development, no more complex projects have ever been conceived than Prometheus in Europe and IVHS in the United States. Prometheus stands for Program of European Traffic with Highest Efficiency and Unprecedented Safety, while IVHS is short for the Intelligent Vehicle-Highway System. The public sector at state, national, and international levels as well as industry, academia, and consumer groups continue to advance these programs, which present unprecedented challenges for cooperation in very complex networks of communication and coordination.

The strategic plan for IVHS in the US[1] gives us an impression of this unique scenario. Although IVHS as a consolidated program is only two years old, already more than 50 operational test sites are in place, and the projected expenditures for IVHS deployment in the US run beyond $200 billion over a 20-year period.

Prometheus was conceived in 1986 as a joint precompetitive research and development program by the European automotive industry in five countries: France, Germany, Great Britain, Italy, and Sweden. It now involves 18 car companies, many electronics and supplier companies, over 100 research institutes and universities as well as numerous consulting companies and public authorities such as those for transportation and telecommunications. In spite of its significance, the annual Prometheus budgets of about $100 million have been lean, with more than two thirds provided by the industry and one third by national ministries of research and technology. Road transport-related programs of the European Community like DRIVE (Dedicated Road Infrastructure for Vehicle Safety in Europe) supplement the effort, and, recently, numerous test sites have been established in Europe with partial regional, national, and European Community support.

In Japan, several major projects are under way: RACS (Road/Automobile Communication System), AMTICS (Advanced Mobile Traffic Information and Communication System), and recently VICS (Vehicle Information and Communication System).

0272-1732/93/0200-0007$03.00 © 1993 IEEE

Table 1. Intelligent Vehicle Highway Systems benefits matrix (in percentages).[2]
(Copyright 1992 US Government Printing Office. Reprinted with permission.)

Benefits	Individual travelers	Fleet operators	Businesses	Government agencies	Society at large
Safety	40	20	—	—	40
Congestion	30	20	—	20	30
Environmental benefits	—	—	—	—	100
Energy conservation	30	10	—	—	60
Universal mobility and accessibility	70	10	—	20	—
Public transportation	60	—	—	20	20
Economic activity	40	—	40	—	20
Law enforcement	—	—	—	30	70

Source: Sigmund Silber

The scope and the progress of these programs are so multifaceted that I've had to deliberately select a certain topical area to give a somewhat concise view in this magazine of the present state of goals and results.

What are the expected benefits of intelligent vehicle-highway systems? A matrix,[2] reproduced in Table 1, addresses the major issues of safety, congestion, environmental benefits, energy conservation, universal mobility and accessibility, public transportation, and economic activity. Prometheus displays a similar ranking when one considers its major European demonstration projects:[3]

- *Safe driving*
  Vision enhancement
  Proper vehicle operation
  Collision avoidance
- *Traffic flow harmonization*
  Cooperative driving
  Autonomous intelligent cruise control
  Emergency systems
- *Travel and transport management*
  Commercial fleet management
  Dual-mode route guidance
  Travel information services

The structure of IVHS again reflects this pattern with its five subprograms: Advanced Traffic Management Systems (ATMS), Advanced Traveller Information Systems (ATIS), Advanced Vehicle Control Systems (AVCS), Commercial Vehicle Operations (CVO), and Advanced Public Transportation Systems (APTS).

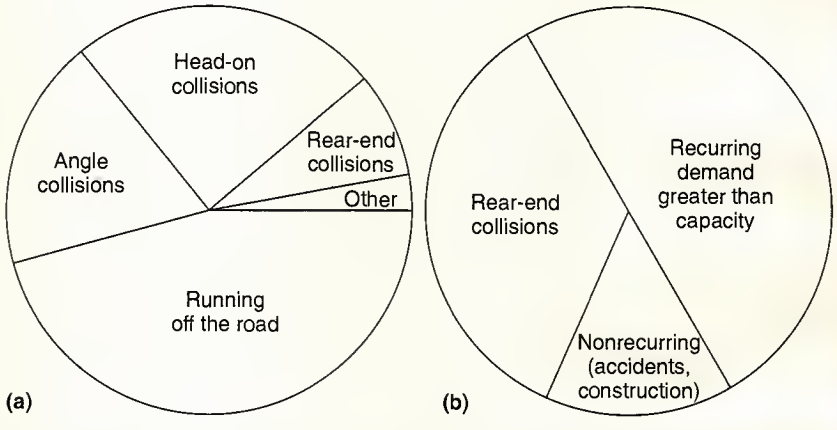


Figure 1. Causes of fatalities (a) and congestion (b).[1] *(Copyright 1992 Intelligent Vehicle Highway Society of America. Reprinted with permission.)*

The potential benefits from the view of the individual driver most likely focus on safety and mobility. The program areas of safe driving and traffic flow harmonization in Prometheus as well as Advanced Vehicle Control Systems in IVHS address these topics most closely. I've selected the articles in this issue of *IEEE Micro* accordingly.

A look at the causes of road traffic accidents and congestion (Figure 1) immediately shows the need and potential for significant improvements through the realization of what we colloquially call the electronic copilot in the car.[4] Over 90 percent

of all accidents in road traffic still result from human error.

Although the human brain's capacity for learning, association, memory, and processing far surpasses any computer conceivable at present, it is decidedly slow. The human reaction and decision cycle takes about 2 seconds, which is equivalent to traveling 50 meters in high-speed road traffic. Delays and errors in braking, passing, negotiating obstacles or curves, or recognizing signs and signals result in a presently unavoidable toll of accidents. Advancing the reaction time by just 1 second would eliminate 80 percent of these accidents.

Fatigue, misjudgment of safety margins, and incomplete knowledge of the status of our own vehicle and of other participants and objects in our relevant road traffic zone are the other major reasons for accidents and congestion. These causes indicate that significant benefits can and will only be possible if the electronic copilot in our car can communicate with other traffic partners, with the roadside, and with the travel management system.

Clearly, this scenario of road traffic differs considerably from what we have today, and it will take the cooperation of all constituencies to move into this new era. However, two major forces may bring about change:

- congestion and pollution approach total deadlock faster than present relief programs can affect, and
- big opportunities exist for the world's advanced economies to serve their citizens in the need and desire for safe individual mobility.

In the first article, "Research and Development Needs for Advanced Vehicle Control Systems," Steven Shladover of the University of California, Berkeley, who is also chair of the IVHS Advanced Vehicle Control Systems committee, identifies what must be accomplished in the new control systems. The second article presents an exemplary realization of an integrated system: the Arena public road test site in West Sweden. Its author, Ulf Palmquist of AB Volvo, is a deputy member of the Prometheus Steering Committee and chair of the Technical Board of the Swedish Road Traffic Informatics Program.

Given this scope of road traffic electronics, it is evident that mainstream microelectronics will not directly qualify for the car control functions, which are all safety-relevant. Car control electronics must have

- avionics reliability,
- no box protecting it from the environment,
- small volume and weight like a pocket computer, and
- lower cost than individual consumer electronics.

Among all these design and manufacturing challenges, microelectronics reliability is most important. Accordingly, reliability research has been a common thread in the Prometheus PRO-CHIP (Prometheus Custom Hardware for Intelligent Processing) subprogram, a basic research program in which over 40 institutes in France, Germany, Italy, and Sweden participated. Enrico Zanoni of the University of Padua, Italy, who has been the European lead researcher on reliability in PRO-CHIP and who has also been instrumental in establishing the reliability laboratory at the national institute CSATA, Bari, Italy, summarizes these activities in his article, "Improving Reliability and Safety of Automotive Electronics."

Advanced vehicle control systems will benefit from any imaginable development of new hardware and software with a special quest for robustness and cost. I've chosen two examples to indicate feasible solutions. Vision enhancement in fast-changing traffic scenes is possible with a high dynamic-range, random-access silicon camera. This is a prerequisite in a system for longitudinal and lateral car control. Given that support, it is still an intricate task to mimic the steering behavior of an alert driver. The concluding article describes a trained digital neurocontroller that serves as the steering assistant in a Mercedes car, which is under continuous test in normal road traffic.

Any view of car control systems presently under development or test should conclude with the comment that the deployment of these systems will be characterized by three stages to be accomplished over the next 20 years:

- advice and warning systems,
- support systems, and
- control systems.

IN THE SPIRIT OF THE UNIQUE COOPERATION in electronic road traffic as a significant civil technology research and development program, I must thank the many experts for their support. Special thanks go to the authors and the reviewers of the articles in this issue. I gratefully acknowledge the members of the European Steering Committee of PRO-CHIP and their contributions. They represent the many helpful scientists in Prometheus: Gianni Conte, Parma, Italy; Daniel Estève, Toulouse, France; and Peter Weissglas, Stockholm, Sweden.

References

1. *Strategy Plan for Intelligent Vehicle-Highway Systems in the United States*, IVHS America, 1776 Massachusetts Ave. NW, Suite 510, Washington, DC 20036, May 1992.
2. "Public and Private Sector Roles in Intelligent Vehicle-Highway Systems (IVHS/Deployment)," *Proc. Seminar Federal Highway Administration*, No. 1992-625-205/60497, US Government Printing Office, Washington, D.C., Apr. 1992.

Call for Articles

IEEE Micro plans a special issue on standards to appear in December 1993. Particular areas of interest include the following, although related topics are welcome:

- Impact of standards on technology adoption
- Open systems and standards
- Government involvement in standards
- Legal factors in standards
- User demand for standards
- International factors in standards development
- Economic costs and benefits of developing and using standards
- Consortia, user groups, and other non-SDO standards developers
- Standards development process
- Evolution of standards
- Measurement of standards quality

Interested authors should submit an abstract of 500 words or less by **March 15, 1993**, to the Guest Editor (e-mail is preferred):

Stephen L. Diamond
SunSoft, Inc.
2550 Garcia Ave., M/S MTV08-221
Mountain View, CA 94043
Phone: (415) 336-4190
Fax: (415) 336-4477
steve.diamond@eng.sun.com

Authors needing guidelines for submission of articles should contact:

Claire Weller
IEEE Micro
PO Box 3014
Los Alamitos, CA 90720-1264
Phone: (714) 821-8380
Fax: (714) 821-4010

3. "Prometheus 'Safe and Effective Mobility in Euorpe'," Prometheus Office, Postfach 600 202, D-7000 Stuttgart 60, Germany, Sept. 1991.
4. B. Hoefflinger, "Computer-Assisted Motoring," in *Scientific Europe*, N. Calder, ed., Nature & Technology Scientific Publishers Ltd., Maastricht, Brüssels, and London, 1990, ISBN 90-73035-06-6/CIP, pp. 28-33.

Bernd Hoefflinger is director of the Institute for Microelectronics Stuttgart, which develops, manufactures, tests, and qualifies application-specific microchips for industrial applications. He is also in charge of the electronics manufacturing program at the University of Stuttgart. Earlier, he worked as a scientific staff member and as the founder of the MOS Integrated Circuits Division for Siemens AG, Munich. He has held faculty positions with the Electrical Engineering Departments at Cornell University, Ithaca, New York; the University of Dortmund, Germany, as founder and head; the University of California, Berkeley, on a sabbatical; and the University of Minnesota and Purdue University, Indiana, as head. For the Prometheus Eureka Research Program, he serves as the European coordinator of the basic research program, "Custom Hardware for Intelligent Processing" (PRO-CHIP).

Hoefflinger received his diploma degree in physics from the University of Göttingen and his PhD degree from the Technical University of Munich.

Direct questions regarding this article to the author at the Institute for Microelectronics, Allmandring 30A, 7000 Stuttgart-80, Germany; or fax to +49 711 685 5930.

Reader Interest Survey

Indicate your interest in this article by circling the appropriate number on the Reader Service Card.

Low 150 Medium 151 High 152

Research and Development Needs for Advanced Vehicle Control Systems

The Advanced Vehicle Control Systems Committee of the Intelligent Vehicle Highway Society of America has identified research and development activities necessary to improve the performance of the surface transportation system. AVCS represent the application of sensors, computers, and electromechanical actuators to provide drivers with warnings of hazards, assistance in controlling their vehicles, or fully automated control of vehicle motions.

Steven E. Shladover

Partners for Advanced Transit and Highways (PATH), University of California, Berkeley

During the middle 1980s, transportation planners and researchers realized that the rapidly worsening problems of the road transportation system would not be addressed adequately, much less solved, by continued reliance on conventional technologies. This realization grew separately among public agency officials, automotive industry managers, and academic researchers in Europe, North America, and Japan. The interested parties on each continent organized themselves to conduct research, development, and demonstration programs on somewhat different time scales and with somewhat different emphases. These activities have come to be known variously as Intelligent Vehicle-Highway Systems (IVHS) in North America and Road Transport Informatics or Advanced Transport Telematics in Europe. This jargon is not particularly helpful to understanding. A more appropriate term would simply be "Intelligent Transportation Systems."

The unifying themes among these activities are the application of information technologies to the operation of road transportation systems in a much broader fashion than ever before, and the integration of travelers, vehicles, and roadway infrastructure into a comprehensive system by use of the newly available information. Such applications of information technology are relatively commonplace in the air, rail, and marine transportation domains today. However they are extremely rare in road transportation, despite the dominant role that rubber-tired transport maintains throughout the industrialized world.

In North America an ad hoc group of academic, government, and industry people, who met periodically from 1988 to 1990 under the name of Mobility 2000, defined the basic outlines of IVHS. This effort began with a group of about 40 people meeting at the University of California, Berkeley, in March 1988 and concluded two years later with a meeting attended by several hundred participants in Dallas. Mobility 2000 was succeeded by a more formal organization called the Intelligent Vehicle Highway Society of America (IVHS America), which was chartered in 1990. This group has prepared a strategic plan for the development and deployment of IVHS in the US, which has been followed by more specific near-term program recommendations to the US Department of Transportation.

The goals of the IVHS program are to improve the performance of the surface transportation system in a wide variety of dimensions by

- reducing traffic congestion;
- improving safety;
- enhancing mobility of travelers, especially the elderly and disabled;
- increasing the productivity of the transportation infrastructure;
- reducing energy use;

0272-1732/93/0200-0011$03.00 © 1993 IEEE

AVCS can provide warnings to the driver, assist in controlling the car, and even take complete control of the car's movements.

- reducing pollution;
- reducing capital and operating costs;
- increasing the viability of public transportation;
- responding more effectively to incidents; and
- increasing the ease and convenience of travel.

These goals should all be promoted by the use of IVHS technologies.

Inherent to the concept of IVHS is the use of information to link the traveler, vehicle, and roadway infrastructure as an integrated system. This means that new organizational and managerial approaches will be necessary to lead to deployment and operation. The technological linkages cannot be accomplished unless the private developers of vehicles and in-vehicle technology, the public owners and operators of the roadway, and the commercial and individual travelers work together to decide what they need and how to achieve it. The political, organizational, and managerial efforts associated with this coordination across sectors are likely to be as challenging as the technology development efforts needed to bring IVHS forward to deployment.

The IVHS program in the US has been subdivided into six functional areas, three of which are oriented toward the following families of technology: Advanced Traffic Management Systems (ATMS), Advanced Traveler Information Systems (ATIS), and Advanced Vehicle Control Systems (AVCS).

Three functional areas are oriented toward application domains: Commercial Vehicle Operations (CVO), Advanced Public Transportation Systems (APTS), and Advanced Rural Transportation Systems (ARTS).

A technical committee in IVHS America represents each of these functional areas, with cross-cutting committees in a variety of other areas:

- Systems Architecture,
- Safety and Human Factors,
- Standards and Protocols,
- Institutional Issues,
- Legal Issues, and
- Benefits, Evaluation, and Costs.

IVHS, including its most advanced element, AVCS, is by no means a creation of the most recent decade. The concept of automating traffic flows was portrayed as part of the General Motors Futurama exhibit at the 1939-40 New York World's Fair. General Motors and RCA tested some of the technology of vehicle control on experimental vehicles in the 1950s and 1960s,[1] and analogous experiments were also conducted in Japan[2] and England[3] prior to 1970. Ohio State University conducted an extended program of automated highway research in the 1960s and 1970s under the leadership of Robert Fenton.[4]

In the late 1960s and 1970s, the interest in automatic control of rubber-tired vehicles shifted from the application on private passenger cars to transit operations on exclusive guideways, known as Personal Rapid Transit (PRT) or Automated Guideway Transit (AGT).[5-8] Hybrid automated vehicles, capable of operation both on guideways and conventional roads, became known as Dual Mode.[9] Research results obtained on all of these developments are scattered widely throughout the technical literature, with the heaviest concentrations of papers in the conference proceedings just cited. The *IEEE Transactions on Vehicular Technology* published three feature issues highlighting IVHS and AVCS technologies, scattered at about 10-year intervals.[10-12]

In the present-day IVHS program, the strongest emphasis has been placed on the nearer term technologies of ATMS and ATIS, with considerably less attention having been paid to AVCS. This emphasis is reflected in the principal IVHS conference proceedings of the past several years,[13-18] which have very few if any papers about AVCS. Some of the current AVCS technology research has been reported in a handful of sessions at the three most recent American Control Conferences.[19-21]

We can now focus on the AVCS and the technical issues that the AVCS Committee has identified as needing attention.

Advanced Vehicle Control Systems

AVCS represents a broad grouping of technologies and potential products, not all of which are control systems. This category includes not only systems that can take complete control of the movements of a vehicle but also systems that can assist a driver in controlling the vehicle and systems that provide "high-bandwidth" information to the driver, particularly about imminent hazards. AVCS therefore subdivide into three separate stages of development, which are expected to follow increasingly long (but still somewhat overlapping) development paths:

- driver warning and perceptual enhancement systems,
- driver control assistance systems, and
- fully automated vehicle control systems.

At each stage, AVCS involve interactions among different

vehicles or between vehicles and the roadway infrastructure. The fully automated vehicle control systems, such as the automated highway systems (AHS), are particularly controversial because of their significant difference from present-day operations. Opinion within the IVHS research community and the larger transportation community differs regarding the feasibility, desirability, and time scale for their development and deployment. While some observers concentrate on the potentially very large benefits in safety, capacity, and efficiency that AHS could offer, others concentrate on the technical and institutional risks to overcome and the up-front investments that will be needed to realize those benefits.

Many enabling technologies will be applicable to each of the three stages of AVCS development, and should therefore not be assigned to any one of the three individually. Each stage will have its own target products that will be made available to the public for use. Some of these individual products can be combined to produce more comprehensive systems, with a wider range of public and private benefits. The AVCS subject area has been subdivided according to each of these three dimensions (enabling technologies, target products, and systems) for study. Different kinds of activity need to be associated with each.

The activities needed for enabling technologies include

- definition of performance requirements,
- identification and evaluation of promising existing technologies,
- identification of "gaps" in available technologies,
- basic research and development on needed technologies, and
- adaptation of existing technologies to AVCS needs.

The target products need

- definition of performance requirements;
- selection of enabling technologies to use;
- product design, development, testing, and marketing.

The systems will need

- definition of performance requirements,
- concept design and analysis,
- selection of target products to incorporate,
- research and design of system architecture, and
- coordination of public and private sector roles.

Here, the principal focus is on the enabling technologies, which can serve as the building blocks for development of the products and systems. These are also likely to be more familiar to readers who are not yet well versed in the subject of IVHS. Later I discuss briefly the products and systems in which these technologies will be used.

It is essential to recognize the strong constraints under which these technologies must be brought to maturity for successful use in AVCS.

Enabling technologies for AVCS—Constraints

Subdividing the enabling technologies for AVCS into several common groups eases discussion. These groupings are not entirely distinct from each other, but must be related.

The hardware technologies are generally not very exotic by standards normally encountered in the aerospace, defense, or computer industries. However, it is essential to recognize the strong constraints under which these technologies must be brought to maturity for successful use in AVCS. These are primarily cost, reliability, fault tolerance, and environmental hardening, combined with the basic performance requirements.

Cost. The automotive world is extremely price sensitive. Automotive OEMs take pains to squeeze every penny of avoidable cost out of a vehicle or option, and every dollar of additional unit cost requires major justification. Complete AVCS must be sellable to the end user for several hundred dollars, and probably an absolute maximum in the range of $1,000, according to the currently accepted thinking within the US automotive industry. This factor imposes much more severe unit cost constraints than the aerospace or defense industries are accustomed to. If suitable technological approaches are considered from the start, significant production economies of scale should be expected when yearly sales are in the hundreds of thousands or millions. However, the unit costs and mass production volumes must be considered carefully right from the start.

Reliability and fault tolerance. All AVCS devices have significant safety implications for the equipped vehicles, their occupants, and their neighbors. If they malfunction, they can easily produce accidents, with property damage, injuries, and even fatalities. The existing road transportation system, even with its unacceptably high accident rate, is actually characterized by remarkably high mean times between fatalities and injuries. Recent US traffic accident statistics indicate a mean time between fatalities (MTBF) on the order of one million vehicle hours for all classes of roads, and even higher than that for limited-access freeways. Even if a failure is taken to represent an injury-producing accident, the MTBF is still on

The largest and most important single category is sensors to detect the condition of a car and its driver and its location relative to the roadway and other vehicles.

the order of tens of thousands of vehicle hours. These are remarkably high reliability levels to be achieved by complex technologies. Since one of the primary IVHS goals is improving safety, it will be necessary for AVCS devices to exceed these current effective reliability levels.

The need for very high effective MTBF in the complete system indicates the need for both high reliability and fault tolerance in component and system designs. This factor has implications for both hardware and software designs. It also reinforces the need for extremely low unit costs of components so that the most critical ones can be used redundantly or with voting (selection of majority sensor readings) to enhance system reliability.

Environmental hardening. The environment in which automotive equipment must operate is quite inhospitable. It includes wide ranges of temperature and humidity, substantial noise (acoustic and electromagnetic), vibration, as well as dust, dirt, snow, ice, fog, and other adverse weather conditions. Because of the safety-critical character of much of the AVCS equipment, it really must be able to operate effectively under all possible combinations of adverse environmental conditions, probably even up to a nearby lightning strike, but stopping just short of thermonuclear war or a major hurricane or tornado.

Needed technologies

The enabling technologies for AVCS have been subdivided into categories of sensors, communication, computation, electromechanical actuators, software and systems technologies, and special tools and facilities.

Sensors. The largest and most important single category of needed enabling technologies is sensors to detect the condition of the vehicle and its driver, as well as its location relative to the roadway and other vehicles. The following kinds of sensors are likely to be needed:

- *Ranging devices to detect the spacing and velocity difference between a vehicle and its neighbors, both fore, aft, and to the sides.* The required range is likely to be between 1 and 100 meters, with an accuracy of 1 percent, a sampling rate of at least 20 Hz, and the ability to operate under all weather conditions.
- *Obstacle detection to find hazards in the vicinity of a vehicle so that accidents can be avoided.* These sensors share some of the requirements of the ranging devices but must also be able to distinguish objects other than vehicles. The objects could be people, animals, dropped loads, and other objects sufficiently massive to cause damage to the vehicle if they are hit. On the other hand, the range accuracy needed for this function is probably significantly less than that required for the ranging used in vehicle-following control.
- *Lane sensing to detect the lateral position of a vehicle relative to the center of the lane.* The required range is likely to be up to one full lane width, with an accuracy of 1 cm for small deviations and perhaps 10 cm for large deviations.
- *Vision enhancement to produce an image of the environment ahead of a vehicle.* These sensors enable drivers to see obstacles, other vehicles, their own position in the lane, or any other pertinent items that they would otherwise be unable to see because of darkness, glare, dust, or precipitation. The sensor system must have a range of a few hundred meters under all environmental conditions, with high enough resolution to pick up all relevant hazards. They must also be combined with a compatible display to supply the image to the driver with sufficient resolution, contrast, and brightness.
- *Road friction sensing to measure in real time the coefficient of friction between the tires and the road surface.* The vehicle control systems can then respond appropriately to rapid changes in road conditions (snow, ice, standing water, sand, oil).
- *Absolute location sensing to detect the location of a vehicle along its path, relative to entry and exit points or other mileposts.* If this is to be used only for routing purposes, the accuracy could probably be 10 meters. However, if used to determine locations of vehicles relative to each other for regulating maneuvers, the accuracy will need to be better than 1 meter.
- *Absolute velocity vector of the vehicle, not sensitive to tire slip or loss of traction.* This system would determine magnitude and direction, so that longitudinal and lateral components of motion can be distinguished.
- *Accelerometers to accurately measure (perhaps 1-percent errors) vehicle longitudinal and lateral accelerations, compensated for road geometry effects such as grades and superelevations.* These measurements are needed to enhance the performance of the vehicle control systems and to provide redundancy for other measurements of the vehicle state.

- *Angular rotation rate to measure yaw rate in particular, a very useful measurement for vehicle lateral control.*
- *Linear displacement measurements of suspension deflections and steering system motions to verify that the vehicle responds to commands in the correct way.*
- *Driver performance to identify the alertness of drivers and their ability to control the vehicle safely.* This has two different uses, one to provide a warning to drivers if their performance is degrading while driving and the other to verify the readiness of drivers to resume manual control after the vehicle has been operating under fully automatic control.

In fully automated systems, the vehicle-roadway communications are also vital for system management, routing, and scheduling functions.

Communication devices. This category includes vehicle-vehicle and vehicle-roadway communications.

Vehicles can alert their neighbors within the same and adjacent lanes through short-range, line-of-sight, two-way, full-duplex communications. These communications are needed for coordinated control and maneuvering and to warn of immediate dangers such as obstacles or vehicle failures. They need to be relatively fast, with high bandwidth and extremely high reliability under all conditions.

Vehicles can also use two-way, short- to medium-range communications between themselves and the roadway. Depending on the system design and operating concepts, these may require any of a wide range of capabilities. In particular if these are substituted for any of the vehicle-vehicle communication needs, the requirements will be substantially more demanding than they would otherwise be. Regardless of the use of vehicle-vehicle communications, this function is still needed for supplying system-level control information to vehicles and for notifying the system of any problems that occur on board the vehicles, as well as for passing information between vehicles that are out of each others' sight or communication range. In fully automated systems, the vehicle-roadway communications are also vital for system management, routing, and scheduling functions.

Computational devices. All of the vehicle control functions require processing of sensor data and calculation of control actions (commands to actuators or driver displays). These devices can require a wide range of computational capability. The performance requirements are therefore more uncertain than any of the other enabling technology requirements. For example, if machine vision is chosen as the preferred sensing mechanism for some functions, the computational requirements are likely to be significantly greater than they would be for alternative sensors. The primary issues remain high reliability, low cost, and robustness in all environmental conditions.

Electromechanical actuators. The control assistance and fully automated AVCS functions require means for implementing the control actions, to change the speed or direction of motion of the vehicle. This involves actuation of the engine (throttle), brakes, and steering system. Some of the enabling technologies are already in use on present-day vehicles. These can range from the antilock braking systems that are now widely available to the traction control and four-wheel steering systems that are only available on a relatively few sophisticated automobiles.

Electronic braking control involves full-authority control of the braking effort, ranging from no braking to full emergency braking, with very fast response. This function extends beyond antilock braking, which can only modulate the braking effort initiated by the driver. While the control signals would be electronic, the actual braking effort would probably be hydraulic, under control of an electrohydraulic servo valve.

Electronic engine control involves full-authority control of engine throttle and fuel injection, with very fast and accurate response to changes in commanded engine torque or speed. This function extends beyond traction control, which can only modulate the engine commands initiated by the driver. It is only available today at very high cost on a limited selection of automobiles, in which the driver's accelerator pedal commands are translated into electronic commands to the engine.

Electronic steering control involves full-authority control of the steering angle, with fast and accurate response to commanded steering changes. While the steering control signals would be electronic, electric or hydraulic actuation systems could turn the wheels. Limited subsets of this capability steer the rear wheels of a few current automobiles that offer four-wheel steering.

Software and systems technologies. Even when the basic hardware is available to meet some of the needs of AVCS, software must still be developed so that the hardware functions as needed. This is likely to be the most labor-intensive part of the development activities, as well as the most heterogeneous. Some of the work occurs at the microscopic level within the system, while other work ranges all the way up to the most macroscopic level.

- *Reliable, fault-tolerant system designs.* The combination of hardware and software to produce highly reliable and

Even when the basic hardware is available to meet some AVCS needs, software must still be developed so the hardware functions as needed.

fault-tolerant systems within tight cost constraints will be one of the most challenging topics in all of IVHS. Although substantial effort has been devoted to design of highly reliable and fault-tolerant systems in the aerospace, nuclear, computer, and process control industries, these application domains have not been as cost-intolerant as the automotive-IVHS domain. Extremely high MTBF rates will be needed in complicated electromechanical systems that can be sold for less than $1,000. The cost constraints may produce the need for some significantly new approaches in this arena.

- *Fault detection and accommodation.* All major subsystems within the automobile should have self-diagnostic capabilities, combined with fall-back modes of operation to accommodate faults. While some diagnostics are already being applied on a "static" basis to facilitate troubleshooting by automotive maintenance people, this software will need to be extended to on-line diagnostics, combined with the logic to choose the most appropriate "degraded" mode of operation. Development of these capabilities will require fairly basic work on fault-detection logic, combined with very practical consideration of the implementation means available on automobiles.
- *Data fusion.* AVCS vehicles will be equipped with many sensors, incorporating substantial redundancy to achieve the reliability and fault tolerance goals. Substantial attention must be paid to the design of the data fusion software. This software will combine the outputs of the various sensors with their different accuracies, error characteristics, and failure modes. When the sensors produce seemingly incompatible outputs, the software will have to define how heavily to weigh the competing information to produce a high-confidence estimate of what is really happening.
- *Threat analysis.* The road environment can be remarkably complicated, particularly if no special measures are taken to simplify it for the benefit of automated vehicles. Thus it will be very challenging for vehicle-mounted sensors to interpret the information they receive so that they can distinguish genuine threats from spurious ones. For example, the sensors will have to identify how threatening an oncoming vehicle is on a curving two-lane rural road: Is it staying in its own lane, or is it straying into my lane? It can also mean predicting whether a vehicle crossing in front of my vehicle is likely to collide with my vehicle, or whether the animal on the road in front of me is a bird that can fly away before I hit it, my neighbor's cat, which I should try to avoid hitting, or a squirrel, which I may not mind hitting. These examples are specific cases, which will each require its own logic. This topic is likely to be complicated precisely because of the large number of such examples that will need to be considered.
- *Nonlinear and adaptive control design.* Automotive vehicles are highly nonlinear, and their precise performance characteristics depend on many difficult-to-predict variables. Therefore, nonlinear and adaptive control systems must control these variables consistently, reliably, and with high performance. The theory for design of such systems is still in its relative infancy. Substantial research will be needed to develop control software that can successfully handle the full range of conditions that each vehicle will encounter throughout its useful life. Included are the normal aging of components and subsystems, substandard maintenance, and substantial variations in loading, as well as variability in the weather and road surface conditions.
- *Human interface designs.* AVCS can substantially change the experience of driving in a variety of ways. Interactions between the driver and the vehicle must be understood thoroughly before AVCS-equipped vehicles are made available for public service. In the case of the driver warning and assistance systems, designers must understand how drivers will react to the different kinds of information and control assistance that will be offered, so that the safety and effectiveness of the system are not compromised by unintended human responses. They must also understand what the drivers like and dislike about various aspects of these systems, so that the systems will be sufficiently attractive for people to want to buy them.

  The human interface issues are somewhat different for the fully automated systems, since these represent even more dramatic departures from present-day driving practices. In this case, designers must understand how drivers respond to relinquishing control of their vehicles to the automatic systems, and what performance or operational characteristics of the automatic systems make them more or less attractive to people. We need to understand how much, and specifically what, information drivers want to receive about the operation of their vehicles when they are driving in the automated

mode. The return of control to the driver at the end of the automated stage of a journey also needs significant attention, particularly to establish how to verify that the driver is indeed sufficiently alert to drive safely.

- *Automatic trip routing and scheduling.* Fully automated driving offers the possibility of automatic routing and scheduling of trips to make optimal use of the automated road network. Substantial software work will be needed to develop and refine the routing and scheduling algorithms. These algorithms should permit the simultaneous optimization of individual vehicle paths and network flows in systems that may contain hundreds of thousands of vehicles at a time.
- *Architecture for system integration.* Each stage in the development of all IVHS functions involves making decisions about the distribution of intelligence within the system. AVCS is no different from the rest of IVHS in this need. Defining the most suitable system architecture is a challenging effort because of the multitude of considerations that must be weighed.

  Depending on how intelligence is allocated among individual vehicles, groups of vehicles, local roadside installations, and a central roadside installation, the communication burdens can vary substantially. The costs of the communication must be weighed against the costs of the information storage and processing elements at each location. Designers must take into consideration as well the need for system-level reliability and fault tolerance. All of this must also take into account the varying possible rates of market penetration of vehicle equipment and installation of roadside equipment, which are financed by different sectors of society. The combination of issues such as these imbues the architecture problem with its richness.

Special tools and facilities. The development of AVCS technologies will require the availability of a substantial amount of data, models, facilities, and vehicles that do not generally exist. The time and resources required for acquisition of these special needs must be taken into account in planning the development of AVCS.

- **Data**. Considerable data about current conditions are needed to provide a solid foundation upon which to build the designs of new AVCS intended to help solve today's problems. These include several different categories of data.

  Accidents. Extensive information about the causes and mechanisms of accidents is needed. The AVCS can be targeted at avoiding the most important and serious types of accidents. In addition, authoritative information about the impacts of accidents on congestion helps in estimating more accurately the benefits of accident reductions.

We need to understand how much, and specifically what, information drivers want to receive about the operation of their vehicles when they are driving in the automated mode.

Vehicle characteristics. Dynamic responses of vehicles, including variations with respect to aging and inadequate maintenance, will allow control systems to be designed to satisfy the full range of needed performance.

Driver characteristics. Comprehensive information about driver responses to the variety of stimuli that can be provided by AVCS warning and assistance systems will allow these stimuli to be selected most appropriately.

Road characteristics. The complete range of road geometry and surface conditions in which the AVCS are expected to operate will be combined with the complete range of weather conditions that must be accommodated.

Component reliabilities. Statistically valid data are needed about the reliabilities of components currently used on automotive vehicles and the components proposed for use in the AVCS.

Traffic flows and demand. Transportation planning data are needed to indicate the level of demand that systems must be designed to service.

- **Models.** Data of the type just indicated must be used to develop models that can predict the performance of AVCS at several different levels, from the driver-vehicle interaction to the operation of a complete regional transportation network. They include

  - driver behavior and driver-vehicle interactions,
  - vehicle dynamic response,
  - transportation networks and traffic flows,
  - benefits evaluations, and
  - protocols for evaluation of experiments and operational tests.

- **Facilities**. Large-scale test facilities are needed to evaluate and then demonstrate the performance of AVCS be-

The enabling technologies are not ends in themselves, but they are the means for implementing products that can be used by travelers.

fore these are sufficiently mature to be used in mixed traffic on public roads. The driver warning, perceptual enhancement, and control assistance systems can probably be tested on existing automotive test facilities, in the same ways that other new automotive systems are tested. However, the fully automated systems will require special facilities, with cooperative infrastructure elements installed in, or adjacent to, the roadway. These special facilities must be of sufficient scale to represent the full range of driving conditions that could be experienced in an automated roadway facility, including multiple lanes of traffic, interchanges with local streets, and freeway-to-freeway interchanges.

- **Test vehicles.** Substantial fleets of test vehicles must be equipped with the AVCS technologies. They will make it possible to accumulate enough vehicle hours of operation to prove satisfactory performance under all reasonable combinations of operating conditions and to prove adequate reliability.

  In addition, demonstration of fully automated operations will require the use of a substantial number of vehicles. The number must be sufficient to prove the absence of undesirable interactions among the automated vehicles and at the same time demonstrate the extremely high travel densities that these systems are intended to achieve. All test vehicles will need to have sufficient instrumentation to record experimental results of interest (especially any abnormal conditions or failures).

AVCS target products

The enabling technologies are not ends in themselves, but they are the means for implementing products that can be used by travelers. Certain target products motivate the development of the enabling technologies.

- Driver warnings and perceptual enhancements include frontal collision warning, side/rear/blind spot/lane change warning, lane departure warning, loss of traction (ice) warning, truck rollover warning, vision enhancement, driver performance monitoring/drowsiness warning, and intersection hazard warning.
- Driver control assists include autonomous intelligent cruise control, collision avoidance (braking and/or steering), lane holding (steering assistance), lane change/merge assist, vehicle shutdown based on driver or vehicle condition, and intersection hazard management.
- Fully automated systems include automated vehicles on special-purpose lanes, automated vehicles on their own freeway network, autonomous automated vehicles, and automatic parking.

The fully automated systems are already "systems" that integrate a variety of different functions. The driver warnings, perceptual enhancements, and control assists can be further integrated using a "driver's associate" or "copilot" to prioritize the information coming from the various sensors and individual subsystems. Then the driver would not be overwhelmed with multiple simultaneous stimuli or instructions.

POTENTIALLY SIGNIFICANT IMPROVEMENTS to road transportation operations could be gained through widespread deployment of Advanced Vehicle Control Systems. These improvements are likely to be most apparent in safety and system capacity. Many technologies need to be integrated carefully to make these systems a reality. The bulk of the required effort is not likely to be on the elemental technologies themselves but on their integration and adaptation to the specific application needs of AVCS.

Efforts in this field must remain strongly focused on finding solutions to transportation problems rather than on developing technology for the sake of technology, which can all too easily degenerate into "solutions looking for problems." Close coordination must be maintained between the basic research community, with its solutions (or possible solutions), and the transportation community, with its problems. Since the cultures of these communities are quite different from each other, substantial good will and effort are needed to bring them together into a mutually productive partnership.

Acknowledgments

Much of this article reflects the deliberations of the IVHS America Advanced Vehicle Control Systems Committee and incorporates contributions from many of its members, which

I gratefully acknowledge. It was prepared under the auspices of the California PATH (Partners for Advanced Transit and Highways) Program of the University of California, in cooperation with the State of California, Business, Transportation and Housing Agency, Department of Transportation, and the United States Department of Transportation.

References

1. K. Gardels, "Automatic Car Controls for Electronic Highways," General Motors Research Laboratories, Warren, Mich., Report GMR-276, 1960.
2. Y. Oshima et al., "Control System for Automatic Automobile Driving," *Proc. IFAC Tokyo Symp. Systems Engineering for Control System Design,* 1965, pp. 345-357.
3. K.H.F. Cardew, "The Automatic Steering of Vehicles—An Experimental System Fitted to a DS Citroen Car," Road Research Laboratory Report RRL LR 340, Crowthorne, England, 1970.
4. Ohio State University, "An Investigation of Highway Automation," Ohio State Univ. Reports EES 276A-6, 1969, and EES 276A-15, 1974; "Fundamental Studies in the Longitudinal Control of Automated Ground Vehicles," Report FHWA-RD-77-28, 1976.
5. J.E. Anderson et al., eds., *Personal Rapid Transit,* Institute of Technology, Univ. of Minnesota, 1972.
6. J.E. Anderson, ed., *Personal Rapid Transit II,* Institute of Technology, Univ. of Minnesota, 1974.
7. D.A. Gary, ed., *Personal Rapid Transit III,* Institute of Technology, Univ. of Minnesota, 1976.
8. "Advanced Transit and Urban Revitalization, An International Dialogue," *Proc. Int'l Conf. Advanced Transit,* Advanced Transit Association, Indianapolis, Ind., May 1978.
9. *Proc. TRB Conf. Dual Mode Transportation,* Transportation Research Board, Washington, D.C., Special Report 170, 1976.
10. *IEEE Trans. Vehicular Technology,* Vol. VT-19, No. 1, Feb. 1970.
11. *IEEE Trans. Vehicular Technology,* Vol. VT-28, No. 1, Feb. 1979.
12. *IEEE Trans. Vehicular Technology,* Vol 40, No. 1, Feb. 1991.
13. D.H.M. Reekie, E.R. Case, and J. Tsai, eds., *Proc. First Vehicle Navigation and Information Systems Conf.*, Toronto, Ontario, 1989.
14. "Advanced Telematics in Road Transport," *Proc. DRIVE Conf.*, Elsevier Publishers, Amsterdam, 1991.
15. *Proc. Vehicle Navigation and Information Systems Conf.*, P-253, Society of Automotive Engineers, Warrendale, Penn., Oct. 1991.
16. *Proc. Vehicle Navigation and Information Systems Conf.*, Society of Automotive Engieners, Sept. 1992.
17. *Proc. 24th ISATA Int'l Symp. Automotive Technology and Automation, Dedicated Conf. on RTI/IVHS,* Automotive Automation Ltd., Croydon, England, May, 1991.
18. "Surface Transportation and the Information Age," *Proc. Second IVHS America Annual Meeting,* May 1992, IVHS America, Washington, D.C., Report IVHS-AMER-92-4, Sept. 1992.
19. *Proc. American Control Conf.*, American Automatic Control Council, Institute of Electrical and Electronics Engineers, Piscataway, N.J., May 1990, pp. 268-303, 782-811.
20. *Proc. American Control Conf.,* IEEE, June 1991, pp. 2539-2572, 3078-3124.
21. *Proc. American Control Conf.*, IEEE, June 1992, pp. 1738-1778.

Steven E. Shladover is the deputy director and AVCS program manager of the California PATH program in the Institute of Transportation Studies of the University of California, Berkeley. He previously worked at Systems Control Technology, Inc., in Palo Alto. His research interests lie in the application of advanced technologies to ground transportation problems.

Shladover holds SB, SM, and ScD degrees in mechanical engineering from MIT and actively participates in the American Society of Mechanical Engineers, Society of Automotive Engineers, and the Transportation Research Board.

Direct any questions on this article to the author at California Partners for Advanced Transit and Highways, Institute of Transportation Studies, University of California, Berkeley, Richmond Field Station, Building 452, Richmond, CA 94804.

Intelligent Cruise Control and Roadside Information

The on-board Autonomous Intelligent Cruise Control system controls a vehicle's speed according to the driver's desire and the speed of and distance to the preceding vehicle. Volvo developed, realized, and tested such a system, with enhancements. This system offers a one-directional short-range system for vehicle-vehicle and roadside-vehicle communication and considerations for recommended speed, limits, and traffic signals. It is potentially a key element in linking and integrating the driver-vehicle-infrastructure in future intelligent transportation systems.

Ulf Palmquist

AB Volvo

An on-board vehicle system designed to control the longitudinal velocity at a driver's set value as well as the velocity of and the distance to a preceding vehicle offers several advantages. Compared to the traditional cruise control system found in many vehicles today, the Autonomous Intelligent Cruise Control, or AICC, system uses this information to adjust the vehicle's velocity to that of the preceding vehicle and keep it at a safe distance. Drivers will appreciate the comfort and safety offered by these extra functions. This system encourages smoother driving and, especially when the controllers are well tuned, reduces fuel consumption and the amount of harmful pollutants expelled into the environment, and better harmonizes traffic since acceleration and braking are also reduced.

Adding short-range vehicle-to-vehicle and roadside-to-vehicle communication to an AICC system lets drivers receive more accurate vehicle and traffic data at an earlier stage. (See Figure 1.) Drivers and their vehicle systems can access information about the status of surrounding traffic and take earlier, appropriate actions.

Systems of this sort are currently under intensive study and development in the Road and Traffic Informatics (RTI) programs in Europe, the United States, and Japan.[1,2]

System description and requirements

Simply described, AICC requires, besides the ordinary vehicle sensors and systems, a target sensor to detect and measure the distances to preceding vehicles. Measurement of the relative velocity is an advantage but not a prerequisite. AICC must contain some intelligence and computing power for the evaluation and interpretation of sensor data, determination of appropriate control actions, and selection of information to the driver. The actual velocity control requires local control systems for accelerating and braking. A simple and sufficient man-machine-interaction unit exchanges commands and information with the driver, and a computer network or bus lets data flow between the hardware units. Since this is a real-time multievent application, real-time multitasking software should be used.

Target sensor. The zone in front of the AICC vehicle, of relevance for its velocity control, is not trivial to define. It depends on the demand one has of the system, the handling properties of the vehicle, the actual road conditions (for example, friction between road and tire, road curvature), and the velocity of the vehicle. Since AICC is designed primarily for country and highway driving (not heavy urban traffic), a target sensor should cover a zone of relevance defined as three

0272-1732/93/0200-0020$03.00 © 1993 IEEE

lanes at a distance of zero to, say, 150 to 300 meters; see Figure 2. With a coverage of three lanes, the lane of the AICC vehicle and the lanes to the left and to the right can be scanned. Scanning the adjacent lanes is necessary as vehicles may be overtaking the AICC vehicle and moving into its lane. The range of 150 to 300 meters strongly depends on the conditions under which the AICC system should operate. The well-known formula

$$d = v \cdot T + v^2/2r$$

expresses the distance d required to stop a vehicle at initial velocity v. The first term, $v \cdot T$, is the distance traveled during the reaction time (pure delay) of the driver and/or the system. The second term, $v^2/2r$, is the braking distance required when applying the retardation value r. As an example, consider the case v = 120 km/h (33.3 meters/s), T = 1.0 second, and r = 2 meters/s^2 (maximum retardation for comfort). The braking distance for this case is 311 meters. Hence, if the AICC system must be able to stop in front of static obstacles from an initial velocity of 120 km/h, the target sensor needs to have a range of more than 300 meters.

From a systems point of view, it is natural to require a target sensor that covers a zone of relevance of 150 to 200 meters. The sensor should be able to detect objects from motorbikes to trucks and to measure the distance and the direction to them. As an advantage, the relative velocity can be measured independently, that is, not constructed as a function of distance measurements.

The sensor should be intelligent enough to filter out background noise and disturbances such as echoes from roadside railings and road signs. An ideal target sensor is one that delivers only the distance, the angle, and the relative velocity to objects like motorbikes, cars, and trucks in the zone of relevance. The sensor must function under clear weather conditions as well as in rain, fog, and snow. Today, scanned or multibeam radar and laser systems appear to be the most promising and reasonable choices to implement these needs.

Signal processing and control unit. The hardware unit is a computer that executes algorithms for signal evaluation, interpretation of traffic, decision and determination of control actions, and choice of information to the driver. The signal processing algorithms use the signals from the target sensor and from vehicle sensors as input (velocity, steering angle, yaw rate). Signal processing reduces the noise level of the signals and estimates the states of the AICC vehicle and all other vehicles detected in the zone of relevance. This means that, among others, the two-dimensional velocities of the vehicles and their relative positions have to be estimated.

The control algorithms use the estimated states produced by the signal processing and the driver's set speed as input. Based on these data, the control algorithm determines the correct restriction for the longitudinal velocity (driver's set speed or a preceding vehicle) and calculates the control actions to be implemented by the actuators. The physical form of the control actions depends on how the AICC system is decomposed. A natural decomposition leads to control actions of either velocity or acceleration (positive and negative) commands.

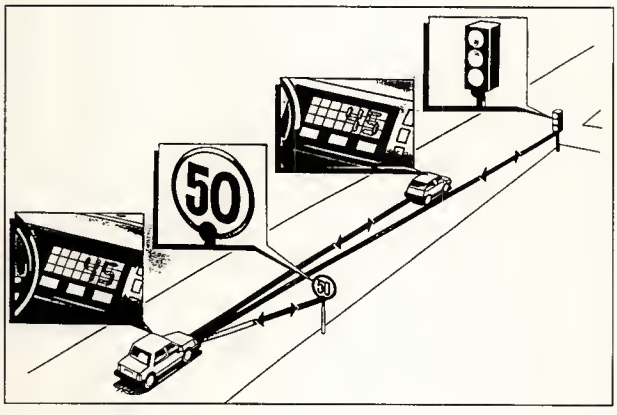


Figure 1. Intelligent cruise control system extended with communications for roadside information.

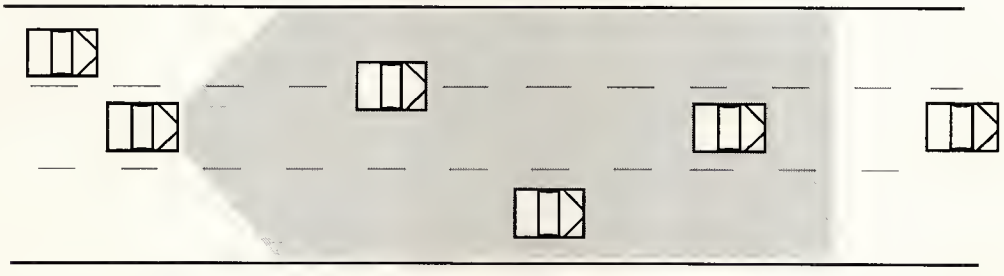


Figure 2. Zone of relevance.

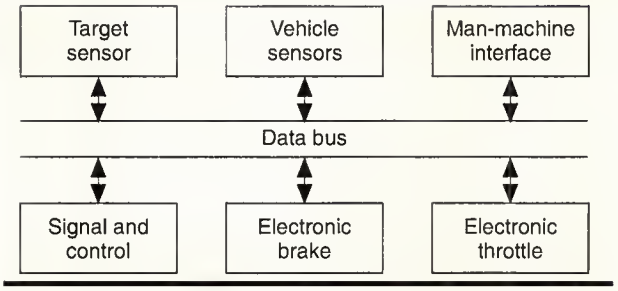


Figure 3. AICC subsystems and their interconnections.

Actuators for velocity control. The AICC system must have local control and actuator systems to adjust the vehicle's velocity in an efficient and smooth fashion. Inputs to these local systems are the control commands from the control unit (velocity, acceleration, and retardation commands).

Two systems are used to fulfill these requirements. An electronic throttle system can accelerate and keep the velocity constant. It can be thought of as a smart actuator that controls the actual acceleration or velocity close to the commanded one. The other is an electronic braking system in the form of a smart actuator that adjusts the actual retardation so it is sufficiently close to the commanded one.

Man-machine-interaction. This unit exchanges commands and information between the driver and the AICC system in a simple way at a sufficient level. The driver should be able to give the following commands and values to the AICC system:

- activate system,
- deactivate system,
- reactivate system,
- set speed value,
- increase set speed, and
- decrease set speed.

This input can, of course, be facilitated in many ways. Say the driver is not allowed to choose combinations of the functions of the AICC system (it is either turned off or turned on with all functions in operation). It seems most reasonable then to use the commonly accepted input pushbuttons of the traditional cruise control system for the AICC system also.

The driver must also be able to override the system manually at any time. Therefore, when wanting to go faster (pressing the accelerator pedal) or slower (pressing the brake pedal), the driver overrides the system. Note that the driver is fully responsible for the vehicle and its operation and, consequently, must have the overall control of it.

At any moment the status of the AICC system and its operation should be clear to the driver. Therefore, it must at least deliver the following information:

- verification of the driver's input,
- mode of operation (passive or active), and
- object for control (driver's set speed or preceding vehicle).

Today, no clear recommendations can be given on the content and form of this information to the driver. Displays and artificial voice are, of course, considered as candidates for output media.

Network and software architecture. The units in the AICC system have computing needs and capacities, and they continuously exchange information and commands. Implementing such a system requires an efficient network and software architecture. The philosophy of system design today is to distribute the tasks, with the corresponding computing capacity, and to connect these local units (or local nodes) with a common data bus. Variables and signals used only within a local unit are restricted to that unit, while variables and signals of relevance to more than one local unit are passed on to the common data bus and accessible to any connected unit. Figure 3 describes this structure.

Each unit must have an interface layer of hardware and software toward the bus to satisfy the specification of the common data bus. Furthermore, global variables and protocols for their transfer have to be defined. The implementation of application software, limited and local to a unit, should be possible with the only restriction that it does not disturb the transfer of the common variables at the data bus.

Short-range communication. By definition, the truly autonomous intelligent cruise control system uses only passively reflected waves from the target sensor in its detection and measurement of preceding vehicles. The advantage of this system is that it does not require equipment mounted on other vehicles for their cooperation. The drawback is that the data received is not always reliable; often the noise level is high, and echoes from objects along the roadside may disturb the measurements and target tracking.

Adding a system for short-range communications, SRC, between vehicles and between the roadside and a vehicle can improve the detection and measurement of preceding vehicles and also extend the functionality of the AICC system. It can transfer absolute or relative positions and vehicle state data from vehicle to vehicle as well as data from roadside equipment, for example, speed limits, status of traffic signals ahead, curvature of bends in the road. With this subsystem, the AICC system can more accurately adjust the vehicle's velocity. The SRC system can be incorporated as just another sensor of velocity restrictions within the AICC. With a structure of the hardware and software as just explained, it is quite easy to include the data from this sensor.

Volvo's AICC system

The AICC system we developed and designed assists drivers in adapting their speeds with regard to

- the desired cruise speed,
- the distance to and the velocity of the preceding vehicle,
- speed recommendations and limits, and
- traffic signals and Green Wave systems. (A Green Wave system constitutes a number of coordinated traffic signals yielding green periods at the arrival of vehicles traveling in compliance with the recommended speed.)

Note that the Volvo AICC system has functions similar to those described in Figure 1.

Though the system structure largely follows that already described, it differs in one major way. Our AICC system is equipped with a transponder-based SRC system for acquisition of data from preceding vehicles and the roadside.

Vehicle. The vehicle equipped with our AICC system (pictured in Figure 4) is a standard 1991 Volvo 960 model with electronic control of the gear box.

Target sensor. The autonomous operation of the AICC system uses a target sensor made by Leica[3] and consisting of five fixed, nonoverlapping infrared beams. Each beam has a range of 150 meters and an angular coverage of 1.5 degrees, horizontally and vertically. Since the beams are not active simultaneously during measurement, it is possible—by knowing which beam caused a received echo and measuring the time of flight—to obtain the distance and angle (crude, in multiples of 1.5 degrees) to the reflecting object. The sensor cannot distinguish between objects separated less than 5 meters longitudinally, and it delivers the measurements corresponding to the closest object (it yields the first and often strongest echo). The sensor can detect objects the size of motorbikes, cars, and trucks. It also easily relays echoes from road signs and other objects along the roadside. Relative velocity is not available from this sensor.

Short-range communication. To transfer data from the preceding vehicle and the roadside, our AICC vehicle uses a transceiver/transponder-based SRC system. The Swedish Institute of Microelectronics developed this system, named Compose, within the Swedish RTI program.[4] (COM stands for communication and POS for position.)

In the AICC vehicle a transceiver unit, mounted in the front, transmits 17.5-GHz microwaves. Any transponders, in the rear of preceding vehicles and as beacons along the roadside, receiving the microwaves amplify the magnitude and modulate the frequency of the waves before reflecting them. The modulation allows the reflected wave to carry data. The transceiver unit measures the phase shift of the reflected wave and its delayed arrival between three patched antennas and determines the distance and angle to the transponder. Therefore, both measurements of the transponder position and data transfer are possible with the Compose system.

The transponder modulates the frequency according to either a programmed static data set in the transponder (static transponder) or data fed into the transponder continuously from an external device (dynamic transponder). The static transponders mainly supply static roadside information, while the dynamic transponders transmit time-variant data, for example, vehicle state data, status of traffic signals, and Green Wave periods.

Figure 4. Volvo's AICC demonstrator vehicle.

In our AICC system, the major task of the Compose system is to obtain speed recommendations and limits, traffic signal status, and other road sign information.

Signal processing and control. We developed and implemented model-based methods for the signal processing of sensor data and decisions and determinations of control actions in the signal processing and control unit of the AICC system.

As described earlier, the signal processing unit takes the target sensor data and—provided, of course, that the vehicle is equipped with a transponder—the data transmitted by the Compose system from the preceding vehicle as input. The signals from the sensors in the AICC vehicle (speed, steering angle) also become inputs. State estimators, constructed from dynamic models of the movement of the AICC vehicle and preceding vehicles, use these inputs to estimate the states of the AICC and preceding vehicles. These extended Kalman filter[5-6] state estimators combined with gating techniques[7] initiate, track, and delete model states of target vehicles.

The control unit has to take into account the following five restrictions or control objectives:

- driver's desired cruise speed,
- distance to and velocity of the preceding vehicle,
- actual speed limit,
- speed limit ahead, and
- traffic signal ahead.

The separate realization of each of these five restrictions is a control problem in itself; some are not at all trivial to fulfill.

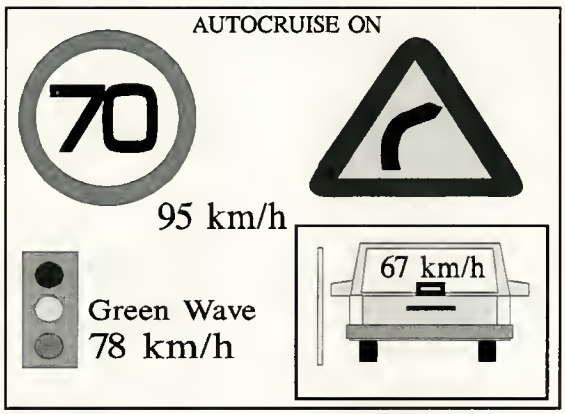


Figure 5. Displayed AICC information.

Furthermore, the combination and simultaneous execution of each requires a well-structured control unit. Note that the first four restrictions imply upper bounds on the velocity and acceleration of the AICC vehicle. The fifth restriction implies an upper and a lower bound on the velocity trajectory of the AICC vehicle so it can pass the traffic signal during a green period.

Since a velocity change of the AICC vehicle can be achieved by an acceleration command to the actuators, it is natural to design separate regulators—one for each of the five restrictions—whose outputs are acceleration commands. The outputs from the first four regulators will be the upper bounds on the permitted acceleration, while the fifth regulator will yield an interval for the acceleration.

Finding the minimal acceleration command among those from the five regulators lets us find the restriction that overrules the other four restrictions and determine which control command should be implemented. This procedure has several advantages. Each regulator can be separately designed to meet the corresponding restriction or criteria. Viewing the acceleration command instead of the actual velocity restriction implies better prediction of how the state will satisfy the restriction. Furthermore, this structure is flexible in the sense that other restrictions can be incorporated in the same fashion, for example, safe driving through a sharp bend with preinformation about the curvature and road/tire friction.

Actuators for velocity control. For the actual control of the longitudinal velocity, we installed two local control and actuator units in the AICC vehicle. These are a throttle system from Hella and a braking system from Bosch; both are electronically controlled.

The throttle system can be operated in three different modes, yielding a choice between the following desired control commands: speed, acceleration, and throttle angle.

The braking system is basically the Bosch ABS model with an electronically controlled plunger system above the ABS level. (The ABS function guarantees antilocking of the brakes. Since the AICC operates above the ABS level, the antilocking function is kept intact.) It can be operated in either of the two control command modes: desired retardation or desired brake pressure.

Man-machine-interaction. The MMI technique in our current AICC system has not been finally developed or adapted to the driver's need and ability. We used the pushbuttons in the traditional cruise control system for the input of the driver's commands and set values. These are

- activate system,
- deactivate system,
- reactivate system (resume),
- set speed value,
- increment set speed, and
- decrement set speed.

When the driver pushes the set button to activate the system, the set speed value is taken as the actual velocity of the vehicle simultaneously. The driver can override the AICC system at any time by pressing the gas pedal or the brake pedal.

Information from the system to the driver is shown as symbols (see Figure 5) on a color display mounted in the dashboard. Basically, the display shows information regarding the four restrictions of actual speed limit, driver's desired cruising speed, traffic signal ahead and its green period, and the distance to and the velocity of a preceding vehicle when they are potentially in force. The display indicates the symbol corresponding to the restriction the control unit has chosen by outlining it in black borders.

This information lets drivers see that the system has interpreted the situation correctly and that it takes the appropriate control actions. The displayed information also allows the system to operate in an informative mode. That is, the AICC system only delivers the information to the driver, who in turn must manually control the velocity of the car. The AICC system, in this mode, does not implement any control actions. We plan to use and examine this mode in the development phase. The upper-right quarter of the display shows road signs that do not necessarily contain information for AICC velocity control but are relevant to the driver for the safe operation of the vehicle. As an example, the upper-right corner of Figure 5 displays the road sign for a sharp bend.

We do not expect nor intend this MMI description to be the one used in a final AICC system; we designed and used it only for the purpose of development.

Network and software architecture. A common controller area network (CAN)[8] carries out the information exchange among the components and subsystems of our AICC system. The software is divided into three layers. The top application layer contains the application programs (signal

processing, controllers), which are implemented in the form of C processes. The next layer provides real-time, multitasking services to the application programs. This layer uses a vehicle distributed executive (VDX), which is an operative software for the organization of distributed real-time processes and the network communication between them.[9] Finally, a network layer connects the VDX to the CAN network. Though use of the various layers has many purposes, one is worth mentioning: The application programmer should not have to bother with multitasking and network services.

Linking to the infrastructure

The AICC is in itself truly an interesting and promising system. However, its value can be further enhanced by using the SRC link to pass information from the infrastructure to the AICC vehicle and its driver. A requirement is, of course, that the roads and streets are equipped with transponders yielding sufficient and necessary information.

The system requires two basic types of information: static and time variable (dynamic). The static information (speed limits, curvature of bends, warnings of school areas) can be preprogrammed into self-contained transponders that only require a power supply. The dynamic information (road conditions, recommended speeds, traffic signal status) is either directly generated by some smart sensors or provided by a local or regional traffic management center. Both have a data link to the appropriate transponders to distribute this information to passing vehicles.

As an example, consider the case of traffic flow control shown in Figure 6. Sensors along or in the street (for example, induction loops) detect and measure passing vehicles. The sensors feed data about the types of vehicles and their speeds to a local traffic manager whose task it is to obtain an efficient and harmonious flow of traffic. The manager continuously evaluates the received data to find the optimal time settings of the traffic signals as well as the suitable velocity to recommend to the vehicles. This manager also handles requests for intersection priority by special types of vehicles (public transport and heavy trucks).

For a harmonized flow of traffic, control of more than just the traffic signals is necessary; preinformation must be given to the vehicles and drivers approaching the intersection so that they can adapt to the active restrictions in due time. The local traffic manager uses the roadside transponders to distribute adequate data to the passing vehicles. This data should contain information about the distance from the transponder to the intersection, the time until the start and end of the next green period, the cycle time of green period, and the length of queuing vehicles. Then the driver can adjust the vehicle's speed to pass the intersection during the green period. Furthermore, for a smooth flow of traffic the transponders should also give preinformation about the status of the traffic signals at the next two or three intersections ahead.

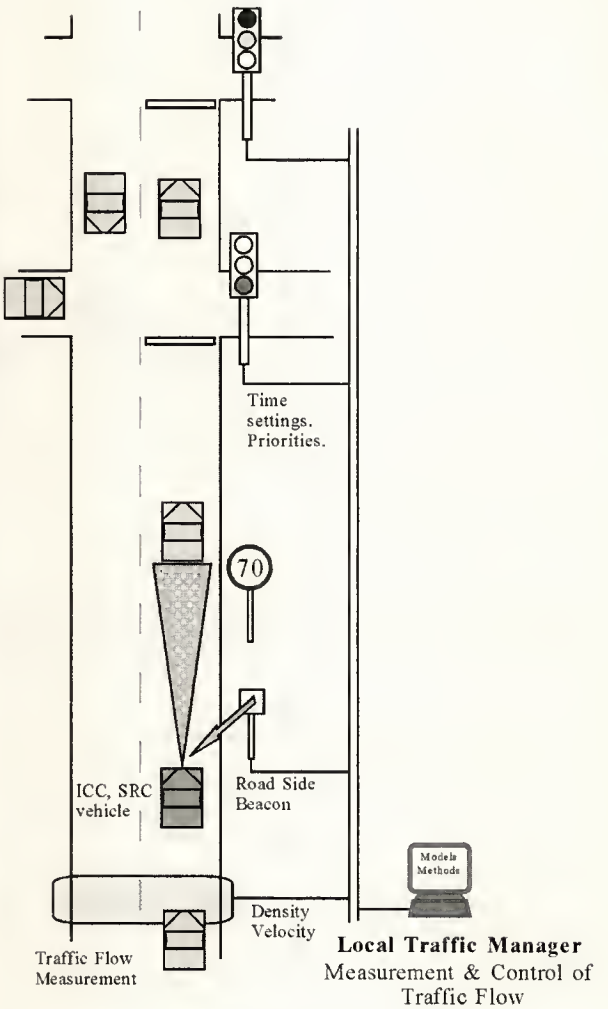


Figure 6. Traffic flow control using intelligent cruise control and SRC.

An AICC vehicle picking up this transponder information determines a velocity profile that also takes into account minimal fuel consumption and pollutant emissions. The display of the complete velocity profile in the form of the recommended velocity lets drivers accept and fulfill a command (informative mode) or choose to feed the data into the AICC system for automatic realization (automatic mode).

Preceding vehicles also have to be considered and evaluated in combination with the traffic signals ahead. The system accomplishes this by continuous evaluation of the information gained by the distance sensor mounted in the front of the AICC vehicle. Consequently, when approaching a vehicle ahead, it changes the priority of the control objec-

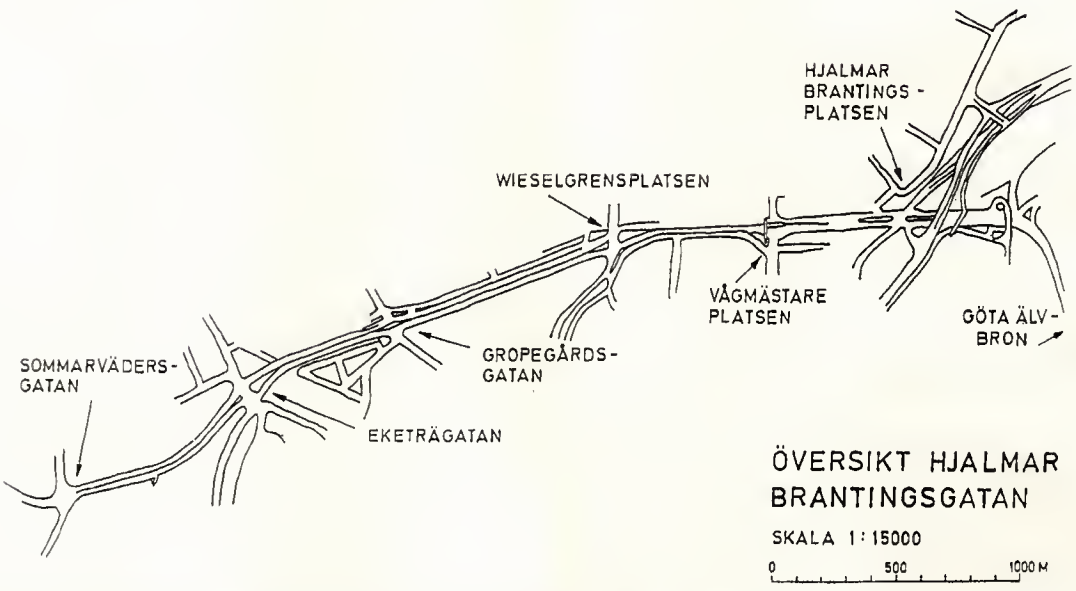


Figure 7. Field trial for traffic flow harmonization at Hjalmar Branting Street.

tive from that of passing at a green period to having a safe distance to the vehicle ahead.

This system can be viewed as a multifeedback control system, in which

- control of the traffic flow is a global loop,
- adjustment of the driver and the AICC vehicle to the traffic signals is an inner loop, and
- adjustment of the driver and the AICC vehicle to the preceding vehicles is a local loop.

These loops constitute the linking of the driver-vehicle-infrastructure.

Note that the idea of giving recommended speed information to the driver is not new. Experiments using radio links to the vehicle have been executed in Wolfsburg, Germany,[10] and Melbourne, Australia.[11] In these experiments the driver was informed of the recommended speed on a display and manually adapted the car's velocity. Reduction in fuel consumption and emission of pollutants, without loss of travel time, were proved.

The Melbourne trials as well as evaluation of systems giving recommended speeds on variable signs along the road[12] show, however, that drivers did not adapt particularly well to the given recommendation. When the recommended speed was low, drivers drove too fast and arrived too early for the green period; when it was high, the drivers drove too slowly. Also when the recommended speed was given only at discrete locations along the road, drivers were not well suited to adapt speeds for any longer distances. Thus, allowing the AICC system to assist drivers and automatically adjust the velocity seems to be a promising step toward harmonized traffic flow.

Field trials

The described RTI technology and systems are not just visions for the future. As a part of the area Driver Assistance and Local Traffic Management within the Swedish RTI program 1991-94, Volvo, Saab, and the Swedish National Road Administration collaborate. They are executing two field trials to explore the technology and feasibility of these systems. These field trials are located in the Arena Test Site West Sweden, which is an open real-traffic RTI laboratory. Located on the west coast of Sweden, the site covers the greater Gothenburg area.

Traffic flow harmonization. Harmonizing the flow of traffic has a potentially positive impact on the protection of the environment and the reduction of fuel consumption. The objectives of the harmonization field trial are to explore and estimate the effects on traffic efficiency, fuel consumption, and pollutant emissions when providing AICC vehicles and drivers with preinformation about speed limits (present and future) and the status of traffic signals ahead.

Hjalmar Branting Street is a highway located in the city of Gothenburg. A 3.5-km stretch of the street has speed limits of 50 km/h and 70 km/h, and six signal-controlled intersections, as shown in Figure 7. The timing of the traffic signals are static but fixed to yield a green period when traveling at

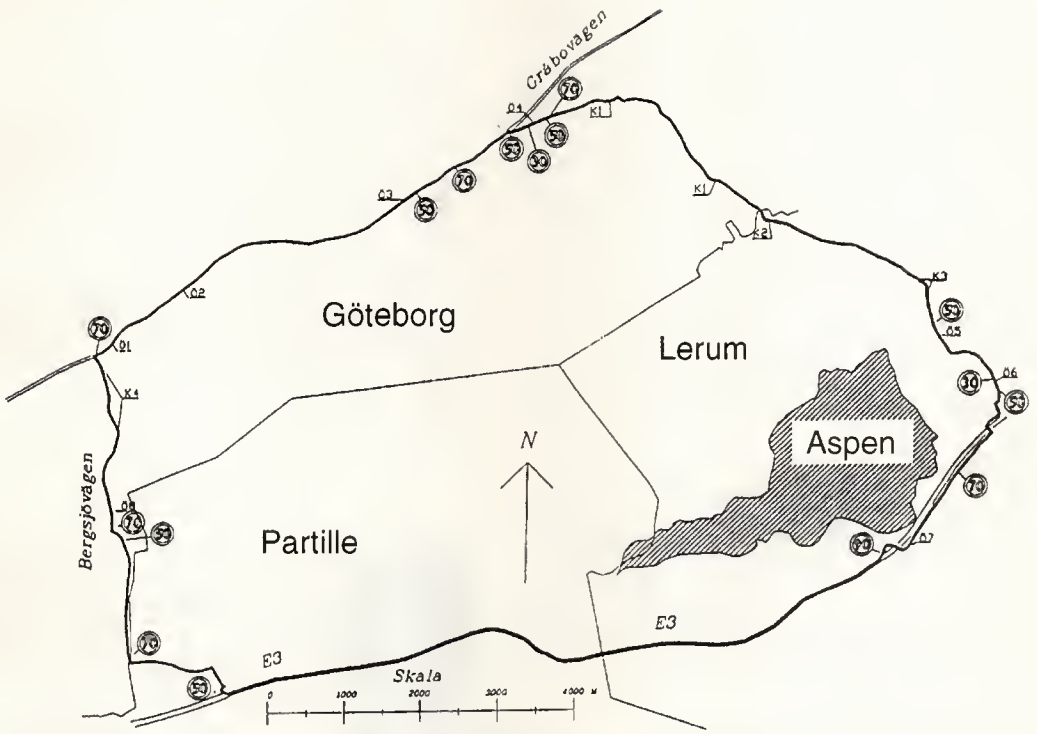


Figure 8. Aspen Track roadside information for active safety.

the appropriate speed. (This speed is not announced and is generally unknown to ordinary drivers.) This street has been equipped with a number of Compose transponders. Each transponder gives information on the speed limit and the status of the traffic signals at the next three intersections, more or less in accordance with that previously described.

AICC SRC-equipped vehicles from Volvo and Saab will be driven in a series of designed runs by ordinary drivers on Hjalmar Branting Street. The following three different driving modes will be investigated:

- *Manual.* With no RTI support, the driver has to manually adjust the speed of the vehicle.
- *Informative.* With information about the speed limit and green period recommended speed, the driver has to manually adjust the speed of the vehicle.
- *Automatic.* With the same information given to the driver as in the informative mode, the AICC system automatically adjusts the speed of the vehicle.

During the test runs variables such as velocity, fuel, and consumption are logged for later analysis of the effects of efficiency, fuel, and pollutant reduction. Based on those results, extrapolations to traffic in larger populations of AICC, SRC-equipped vehicles can be carried out.

Already, this field trial has provided technical experience concerning the SRC link and its advantages and drawbacks in a complex real-traffic environment. One very obvious result in particular is that a real traffic environment demands very robust and reliable SRC systems. More results from the field trial are expected to be available during 1993.

Aspen Track, roadside information for active safety. An SRC link from the roadside to passing vehicles yields the advantage of feeding information into the vehicle system so it can be given to the driver at the correct location and time. This information may change the driver's behavior and, as a consequence, have an impact on the safety not just of the driver but also that of the surrounding traffic and unprotected pedestrians.

East of Gothenburg around Aspen Lake is a track of approximately 35 km of rural and motorway roads, as depicted in Figure 8. Aspen Track has been equipped with transponders transmitting information on speed limits, road curvature, and recommended speeds on sharp bends, warnings of pedestrian crossings, and other relevant information. The field trial explores the effects on driver behavior and safety when using roadside information. The driving behavior of a number of ordinary test drivers using AICC SRC-equipped ve-

hicles from Volvo and Saab when driving on Aspen Track will be studied and logged. The modes of driving are similar to the one used in the traffic flow harmonization experiments.

- *Manual.* With no RTI support, the ordinary driver has to adjust the speed of the vehicle.
- *Informative.* With information about the speed limits, warnings, and so on, the driver has to manually adjust the speed of the vehicle.
- *Assisting.* With the same information, the AICC system automatically realizes the recommended speed.

We executed this field trial at the end of 1992 and expect results from the evaluation in early 1993.

Acknowledgments

The work I've described formed a part of the area Driver Assistance and Local Traffic Management within the Swedish RTI 1991-94 program. Parts of this work are Volvo's contribution to the AICC project within the European Eureka program Prometheus. The members of the Volvo AICC project team are all highly appreciated for their contribution to the development of the AICC system.

References

1. D. Augello et al., "Traffic Flow Harmonization by Cooperative Driving," *ATA Proc. Third Int'l Conf.*, Vol. 3, Apr. 1992, pp. 1889-1898.
2. R.E. Fenton and R.J. Mayhan, "Automated Highway Studies at the Ohio State University—An Overview," *IEEE Trans. Vehicular Technology*, Vol. 40, No. 1, Feb. 1991, pp. 100-113.
3. Leica, "MSAR—Multi Segment Distance Sensor," Tech. Specification Z. 135/88 Rev. B, Leica Heerbrugg Ltd., CH-9435 Heerbrugg, Switzerland, Sept. 1989.
4. P. Henricsson et al., "COMPOSE—A Transceiver/Transponder System for Communication and Relative Positioning," *Proc. Prometheus Workshop*, Prometheus Office, Damiler-Benz AG, Postfach 60 02 02, D-7000 Stuttgart 60, Germany, Oct. 1989, pp. 327-337.
5. R.E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *J. Basic Engineering*, Vol. 82, Ser. D, No. 1, 1960, pp. 35-45.
6. A.H. Jazwinski, *Stochastic Processes and Filtering Theory*, Academic Press, New York, 1970.
7. S. Blackman, "Multiple-Target Tracking with Radar Applications," Artech House, Inc., Norwood, Mass., 1986.
8. D.J. Arnett, "A High Performance Solution for In-Vehicle Networking—Controller Area Network (CAN)," Society of Automotive Engineers, Tech. Paper 870823, Apr. 1987.
9. C. Serrano and M.A. Kung, "VDX: Vehicle Distributed eXecutive," Report 80-NT-442694, Renault, D. R., Paris, France, Dec. 1988.
10. G. Hoffman and W. Zimdahl, "Recommended Speed Indication Within the Motor Vehicle—A Contribution to Fuel Saving Through the 'Wolfsburg Wave' Information System," *Proc. Sym. Energy Efficiency in Land Transport*, Commission of the European Communities, Publ. EUR 12284 EN, Luxembourg, May 1988, pp. 211-222.
11. R.S. Trayford, "Testing Dynamic Advisory Speeds Under Adaptive Signal Control," *Australian Road Research*, Vol. 17, No. 4, Dec. 1987, pp. 243-255.
12. U. Hammarstrom, "Hastighetssignaler—effekt av efterlevnad," [Speed Recommendations—Effect of Compliance], Report 270, Swedish Road & Traffic Research Institute, Linkoping Sweden, (in Swedish), 1981.

Ulf Palmquist is manager of the Systems Engineering Group at AB Volvo, Technological Development, and head of the Volvo AICC project. In the Swedish RTI program, he is a member of the Program Board, chair of the Technical Board, and coordinator of the area Driver Assistance and Local Traffic Management. He is also Volvo's deputy member of the Prometheus Steering Committee.

Palmquist, after graduate studies at Chalmers University Gothenburg, and Imperial College, London, received the MSc degree in engineering physics and chemical engineering. He received an MS degree in process control and control theory from the University of California, Santa Barbara, and his PhD in mathematics and mathematical statistics from Chalmers University.

Direct questions concerning this article to the author at AB Volvo, TD, Systems Engineering, Chalmers Science Park, S-412 88 Gothenburg, Sweden; or fax to +46 31 772 40 86.

Improving the Reliability and Safety of Automotive Electronics

Microelectronics systems designed for automotive applications face an extremely hostile electrical and physical environment. Designers must produce increased component and system reliability while maintaining required compactness and cost-effectiveness levels. Their designs become crucial to all as we devote more electronic systems to safety-critical applications. We summarize the results of the European Prometheus PRO-CHIP research groups working on the reliability and fail-safe operation of microelectronic systems and devices.

Enrico Zanoni

Paolo Pavan

University of Padua

Reliability is a key issue for automotive applications of microelectronic systems. Future cars will be characterized by the increased use of integrated electronic systems (up to 25 percent of the total car manufacturing costs in the year 2000). The functions of these systems will vitally affect car safety and performance. According to a forecast from the European Prometheus project, the automobile of 1995 will have about 100 sensors, 80 actuators, 45 motors, 5 displays, 4 imagers, and 1,000 integrated circuits.[1,2]

The reliability goal forecast by the American automotive industry for the year 2000 is 0.01 ppm cumulative at five years or 50,000 miles, equivalent to 1,800 ignition-on hours. Unfortunately, the current reliability level of electronic components is not yet compatible with the ever-increasing complexity of the electronic systems required for automotive applications in the near future. We need to develop new methodologies for evaluating and improving electronic component and system reliability through research conducted with the close cooperation of system and device manufacturers.

The task is particularly difficult owing to the peculiarities of automotive applications. On one hand the automotive environment is one of the harshest, possibly accelerating a series of different failure mechanisms. On the other hand the required improvement in reliability must be obtained while respecting the specifications of low-cost, high-volume production, light weight, compactness, and short time-to-market imposed by the automotive industry.

We present some of the results and activities of the Prometheus project's PRO-CHIP research groups who studied these problems. We also briefly review the reliability problems most frequently encountered by electronic devices for automotive applications and the procedures the manufacturers use to evaluate and improve reliability of their products.

The automotive environment

Past research of the automotive environment has made its characteristics fairly well known.[2,3] Temperatures within the engine compartment vary from approximately −40 degrees Celsius to +150°C, but the exhaust temperature can be as high as 650°C. Even below the dashboard or within the car interior, the temperature can reach 85°C. Thermal gradients can be extremely high, and a large number of thermal cycles (as high as 40°C/min) is expected during a device's operating life. Thermal cycles promote thermal fatigue phenomena and other failure mechanisms related to the mismatch of the thermal coefficients of the

0272-1732/93/0200-0030$03.00 © 1993 IEEE

Reliability tests for automotive electronic components

The following describes the accelerated testing usually adopted by the automotive industry to evaluate the reliability of electronic components and to qualify new suppliers.

Air-to-air thermal shock. Minimum dwelling time of 18 minutes at each extreme temperature. Common test temperatures are −40°C to 150°C and −40°C to 125°C. Test duration is 1,000 cycles.

Air-to-air thermal cycles. Minimum dwelling time of 15 minutes at each extreme temperature. Minimum ramp rate is 5°C/minute; common test temperatures are −40°C to 150°C and −40°C to 125°C. Test duration is 1,000 cycles.

Liquid-to-liquid thermal shock. Minimum dwelling ingtime of 3 minutes at each extreme temperature. Common test temperatures are −40°C to 125°C. Test duration is 500 cycles.

High humidity/high temperature (optional bias). Components to 85°C and 85 percent relative humidity. Test duration is 1,008 hours.

Life test. Components to their maximum operating temperature and power. Test duration is 1,008 hours.

Hot storage. Components to their maximum operating temperature. Test duration is 1,008 hours.

High-temperature (reverse bias.) Reverse-biased at the device's maximum temperature for 1,008 hours.

Mechanical shock. Three shocks in perpendicular planes. The devices will be shocked at 1,500, 3,000, 4,500, and 6,000 g levels. The minimum acceptable shock requirement will be 3,000 g/0.3 ms.

Vibration. From 5 Hz to 200 Hz for 10 ± 2 minutes per cycle. Repeat the cycle for five hours in each of three planes.

Surge voltage test for capacitors. Voltage surges at 130 percent of rated voltage, cyclically applied 30 seconds on and 30 seconds or 270 seconds off for T_a and electrolytic capacitors. Typical test duration is 1,000 surges.

Intermittent operational life. A 75°C temperature variation for each cycle. Power rated at 85 percent applied for 1 minute and 1 minute cooling for each cycle. Test performed at 0°C or −10°C. Test duration is 20,000 cycles.

Ripple life test. Maximum operating temperature. Bias with 90 percent of rated ripple current and DC bias voltage. Test duration is 1,008 hours.

Autoclave. Requires 121°C, 2 atm, and saturated humidity. Test duration is 96 hours.

employed materials. Relative humidities up to 99 percent, together with the presence of corrosive chemicals (NaCl, CaCl, SO_2, ...) and fuel vapor can accelerate corrosion mechanisms. Instantaneous acceleration and shocks can be as high as 30g.[2]

Severe hazards can be produced from a variety of electromagnetic interference (EMI) and power supply transients, and high-voltage (≈100V, 100 µs to 2 ms) transients result from the presence of large inductive components. Electronic components in the car can be subjected to "load-dump" slow transients. These transients consist of a 10V to 120V positive overvoltage that is superimposed on the nominal 12V supply if a large load or flat battery is disconnected from the electrical system of a vehicle while the engine is running at high speed. This transient can last between 40 ns and 400 ms; it is the most severe electrical overstress that can be induced within the car electrical system.

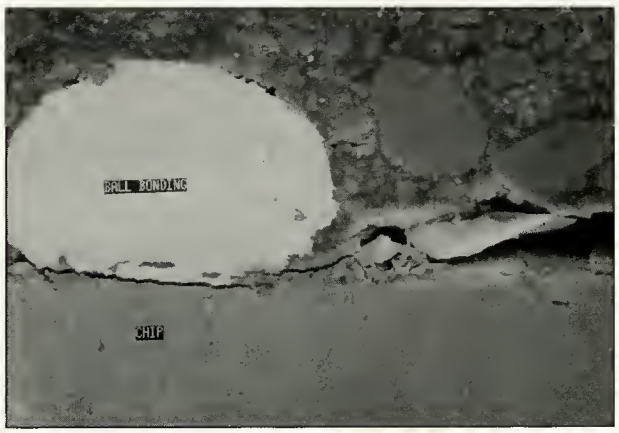
Finally, susceptibility to EMI can be a critical issue due to the presence of intense sources of electromagnetic radiation also within the car itself. In the 5-kHz to 18-GHz frequency range, automotive electronics must withstand field intensities up to 100 V/m without errors.

Reliability testing of electronic components for automotive electronics must assure, in principle, that selected components will meet the failure rate goals specified by system designers. To identify the specific failure mechanisms of electronic components and the related acceleration factors, manufacturers have extensively used accelerated testing. Failure rates in nominal operating conditions have been successfully estimated on the basis of accelerated life test data. (See the box above.)

This traditional approach of "measuring" component reliability by means of accelerated tests and of extrapolating the results to field conditions will become less effective as the failure rates of electronic components continue to decrease. In fact, when the failure rate of the device to be tested is estimated in the 10-FIT (10 failures over 10^9 component-hours) range, the task of evaluating its reliability is cost-prohibitive, both in terms of number of units and in tenns of time.

As a consequence, new ways of evaluating the reliability of electronic devices have been proposed. The "wafer-level reliability" approach consists of highly accelerated wafer-level tests on discrete structures that are designed to address each specific reliability failure mechanism, such as time-dependent breakdown, electromigration, hot-electron effects. However, this method does not cover all failure mechanisms, and even in this case quantitative evaluation of very low failure rates becomes uneconomical.

A detailed in-line control of process variations and of input process variables that may affect device reliability will more effectively "build reliability" into devices. This approach requires the manufacturer of automotive electronic products to



(a)

(b)

(c)

Figure 1. Cross section of a plastic packaged device, showing beginning of delamination of the plastic/chip interface (a,b) and breaking of bonding due to thermomechanical stress (c).

cooperate closely with the IC supplier, possibly developing common programs of reliability assessment and improvement from the very early stage of product design and definition. This does not imply that reliability engineering will reduce itself to a sophisticated kind of process monitoring. On the contrary, a large research effort must be undertaken to better understand device failure mechanisms, to identify which process variables can actually affect the long-term behavior of devices, to develop failure analysis techniques suitable for ULSI (ultralarge scale integration) circuits, and to define design guidelines for reducing the impact of electromigration, "latch-up," electrostatic discharge (ESD), and other failure mechanisms dependent on device scaling. For this reason most of the PRO-CHIP research groups developing new devices or technologies are also involved in reliability characterization, as summarized in Table 1.

Building in reliability

The most frequently observed reliability problems of automotive electronic components generally relate to

1) failure mechanisms due to the package or to the assembling technology;
2) different kinds of electrical overstress, electrostatic discharge, electromagnetic interference; (All these can trigger parasitic bipolar elements of CMOS ICs, that is, latch-up.)
3) breakdown and burnout of power devices; and
4) failure mechanisms accelerated by high temperatures and high current densities.

Failure mechanisms due to thermomechanical stress and thermal fatigue. The trend toward integration of complete systems on a chip requires the placement of larger and larger chips into complex plastic packages with smaller outlines. Unfortunately, the repeated thermal cycling typical of automotive applications can lead to mechanical stress, induced by thermal expansion mismatches between the package materials (plastic compound, silicon chip, lead frame metal). In turn, this induces a series of different failure phenomena. The chip surface and die attach may be subjected to shearing stress, which results in damage to the metallization tracks and passivation cracks; the silicon itself can also be damaged. Say that delamination of the plastic/chip interface occurs, due, for instance, to humidity or contaminants (see the IC cross sections in Figure 1a,b). Then, the plastic can be displaced along the chip surface, resulting in deformations of metallizations and wire bonding and eventually resulting in the breaking of the bonding itself, as shown in Figure 1c.

The formation of a void between the plastic and the chip promotes the penetration of contaminants, inducing the cor-

continued on page 34

Table 1. Reliability research groups within Prometheus PRO-CHIP.

Research title	Institution	In cooperation	Devices studied	Problem or failure mechanism (f.m.)	Failure analysis techniques/ design and test tools
Robust analog design	Institute for Microelectronics Stuttgart (D)	Siemens AG	Current and voltage reference circuits in CMOS; I/O protection circuits	Electrostatic discharge (ESD), EMI	—
Thermal design in power hybrid and high-density assemblies	LAAS CNRS Toulouse (F)	—	Hybrid MOS transistor power switches	Thermal behavior, DC/time dependent	
Reliability of components and assemblies in severe automotive environments	IXL, Univ. of Bordeaux (F)	Siemens Automotive, PSA, Thomson CSF	Tantalum capacitors, ceramic capacitors, packages for SMI	Fail. due to thermal cycles, mechanical shocks, humidity	Impedance f. analy., Piezoel. resonance, acoustic microscopy
Reliability of electronic components for automotive applications	Dip. di Elettronica e Informatica, Univ. Padova (I)	SGS-Thomson Alcatel-Telettra SpA	CMOS ICs, high-, medium-voltage MOS, GaAs devices	ESD, EMI, latch-up, hot-electron effects, f.m. of GaAs devices	Electron microscopy and microanalysis (SIMS, Auger)
Failure mechanisms, functional testing and electron-beam testing	Tecnopolis CSATA, Bari (I)	SGS-Thomson Alcatel-Telettra Siemens Telecom Marelli Autronica	CMOS ICs, high-, medium-voltage MOS, GaAs devices	ESD, EMI, latch-up, hot-electron effects, thermally activated f.m. of GaAs devices	Laser and electron microscopy, emission microscopy
Susceptibility of ICs to electromagnetic interference (EMI)	Dip. di Elettronica, Politecnico di Torino (I)	Centro Richerche Fiat	TTL and CMOS integrated circuits	EMI	—
Functional analysis of intelligent power circuits	LAAS CNRS, Toulouse (F)	PSA, Renault, Motorola Semicon. SGS-Thomson	CMOS, DMOS technology; MOS high-power switch	Latch-up, failure mechanisms of high-power MOS	—
Diagnostic functions for automotive smart-power switches	Robert Bosch GmbH Reutlingen (D)	—	Voltage-protected, smart ignition bipolar driver	—	—
Low-to-high power interfaces	Institute for Microelectronics Stuttgart (D)	Robert Bosch GmbH	Overcurrent, open-load, I/O prot. circuit in CMOS, power MOS	—	—
Smart-power devices in bond and etchback silicon-on-insulator (SOI) technology	Institute for Microelectronics Stuttgart (D)	Daimler Benz AG Forschungszentrum	100V LDMOS, 150V VDMOS on SOI	—	—

D: Germany, F: France, I: Italy

Figure 2. Corrosion of the pad bonding in a plastic device (top), resulting in an open circuit (bottom).

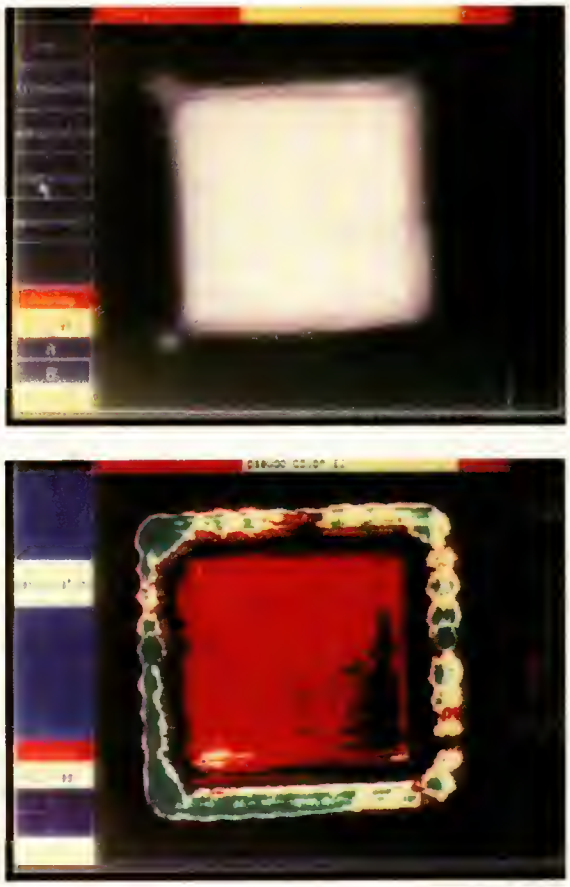


Figure 3. Acoustic microscopy map of a device without defects (top) and of a device in which a void has been created between the plastic package and the chip (bottom).

rosion of the metal pad and eventually resulting in an open circuit, see Figure 2. Defects in plastic encapsulated ICs can be observed nondestructively (without opening the package) by means of acoustic microscopy.[4] Separations of the molding compound from the lead frame or the die are nontransmissive and highly reflective to high-frequency ultrasound. Therefore they appear as high-contrast features in the image. Figure 3 compares the acoustic microscopy map without defects with that of a device in which a void has been created between the plastic package and the chip. The void corresponds to the dark area in the false colors image.

By measuring the linear thermal coefficient of the plastic compounds used for IC packaging as a function of temperature, we can obtain information on possible risks deriving from thermomechanical stresses. At a certain temperature, identified as the glass-transition temperature of the compound, T_g, a remarkable increase in the expansion coefficient occurs. Higher T_g values correspond therefore to increased reliability levels. Figure 4 shows the linear thermal expansion of a package having a T_g = 133°C. This is too close to the operating range of the device and results in bonding deformation after thermal cycling, as shown in Figure 5.

The trend toward increased miniaturization has also resulted in the diffusion of surface-mount technology, and in the need for new substrates that provide a better power dissipation for the components. A PRO-CHIP group directed by Danto at the University of Bordeaux IXL (France) has developed a tool to optimize the thermomechanical behavior of large plastic packages used for surface-mount assemblies.[5] The tool is based on two-dimensional simulation using finite element analysis. The simulations provide information concerning the location and strength of thermomechanical stresses as a function of physical parameters of adopted components.

The authors have submitted different kinds of assemblies,

including plastic quad flat packages (PQFP) and plastic leaded chip carriers (PLCC), mounted on alumina or on isolated metal substrates, to accelerated tests. The tests include –40°C to 150°C thermal cycles, 12 to 60 minutes flat time, and 85°C, 85 percent relative humidity factors. PLCCs on alumina show more failures than PLCCs on metal substrates. The latter provide the best results together with PQFPs on metal. Failures consist of cracks located at the solder-lead interface; these cracks coincide with points of maximum stress as identified by simulations, thus validating the chosen approach.

Another dangerous failure mechanism relates to the thermal fatigue phenomena of power devices, which result from the thermal mismatch between the chip and the header under stresses imposed by temperature or power cycling. The failure mode, for devices using soft solder, usually consists of voids and cracks in the solder material. These defects increase the device's thermal resistance, form "hot spots" in the chip, and eventually induce device burnout owing to thermal instability.

Thermal characterization methods. The channel temperature (T_{ch}) of an electronic device is conventionally described as the sum of the case temperature (T_{case}) and of the product of the power dissipation (P_d) by the thermal resistance (R_{th}). That is, $T_{ch} = T_{case} + P_d R_{th}$. We can evaluate R_{th} and T_{ch} by means of both DC and pulsed electrical methods. These methods are based on the measurement of a device's electrical characteristic (like V_{BE} in a bipolar transistor), which is assumed to depend on temperature according to a known or measured calibration curve.

Electrical methods provide an average measurement of the temperature of the device. Unfortunately, in actual operating conditions, or during accelerated tests, the power dissipated by the device's active areas leads to a nonuniform increase of the device temperature. The T_{ch} value resulting from electrical measurements is therefore an average, weighted in an unknown manner, of the temperature distribution on the device and can therefore be very inaccurate, especially if a small area of high temperature exists within the structure.

The actual temperature distribution on the chip can be measured by liquid crystal techniques or directly observed by means of high lateral resolution infrared (IR) thermography. We can then detect the thermal gradients caused by local differences in the heat dissipation or by structural inhomogeneities. This technique can perform surface temperature measurements of devices with a spatial resolution of 15 µm and a field of view of 1.8 × 1.8 mm². Figure 6 (p. 36) shows the IR thermography map of a 0.25W gallium arsenide (GaAs) MESFET device, biased at P_d = 640 mW, at T_{case} = 24.8°C.

Thermal design of the automotive electronic power circuits markedly influences their reliability; consequently, it is of great importance to develop suitable tools for the thermal design of these circuits. This has been the goal of the project conducted by J.M. Dorkel and collaborators at LAAS CNRS Toulouse, which developed the Pyrtherm package based on

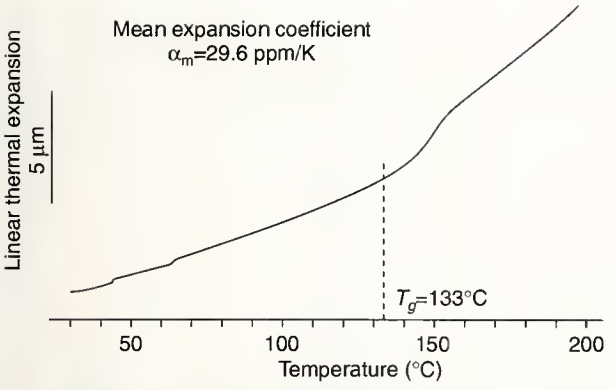


Figure 4. Linear thermal expansion of a plastic package as a function of temperature, identifying a glass transition temperature T_g = 133°C.

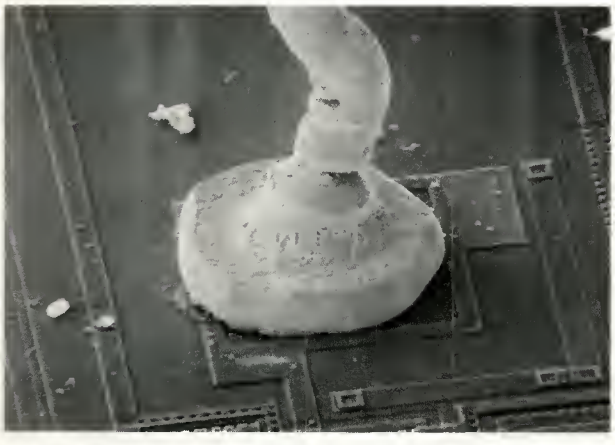


Figure 5. Wire-bonding deformation due to thermomechanical stress in the same IC as in the previous figure.

the use of the Thermal Influence Coefficient.[6,7] This package enables 3D static simulation of complex assembly structures to be easily performed on a personal computer in the interactive mode. It optimizes the thermal structure, hybrid assembly, or component location to obtain a minimal thermal resistance or thermal increase. The group compared the results with temperature maps produced with IR thermography. A 3D thermal step response can be computed for an elementary disk-shaped power source located on top of a rather complicated cooling structure. Using the superposition principle and evaluating a convolution integral, we can compute the thermal response for any power dissipation pulse.

Managing electrical overstress, electrostatic discharge, electromagnetic interference. Because of the intense elec-

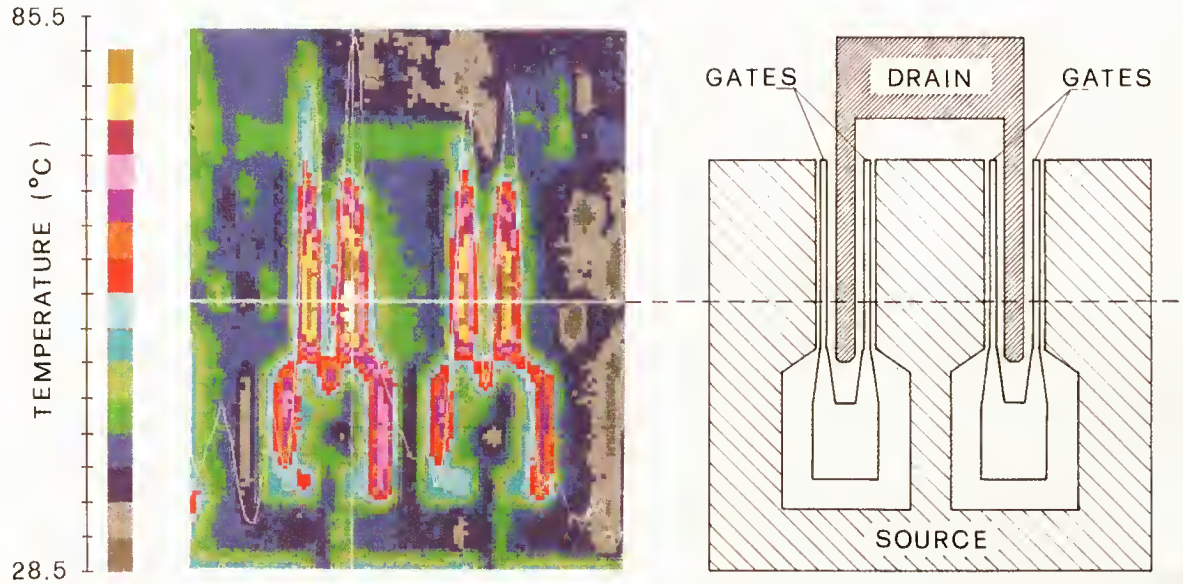


Figure 6. Infrared thermography map of a 0.25W MESFET, biased at P_d = 640 mW, T_{case} = 24.8°C.

trical noise present within a car, failure mechanisms from I/O overcurrent and overvoltages are a serious concern in automotive electronic systems. The techniques for protecting ICs from electrical overstress include both special layout and technology of I/O devices and networks, and specific integrated protection circuits. These circuits enable the device to protect itself against electrical overstress before a permanent failure can occur. More subtle failures can be induced by the triggering of parasitic elements and by electrostatic discharge.

Latch-up in CMOS ICs. Scaling CMOS ICs can include reliability hazards due to the action of the parasitic semiconductor-controlled rectifier (SCR) unavoidably present in bulk CMOS technology. If switched on, this parasitic SCR can connect supply voltage V_{DD} and ground voltage V_{SS} by a low resistance path. This phenomenon is called latch-up and can be induced by overvoltages applied to I/O or supply lines. When latch-up is triggered, the circuit no longer meets its functional specifications, and a large current flows through the parasitic structure, permanently damaging the device.

Figure 7a shows the simplified cross section of a double-well CMOS device. The parasitic PNPN structure consists of NPN (Q_n) and PNP (Q_p) bipolar transistors connected so that one's collector drives the other's base.

The two parasitic transistors Q_n and Q_p are normally in the off state. If one of the two transistors is brought into the on state, and if the current gain product $\beta_n\beta_p$ is sufficiently high, latch-up occurs. Both transistors remain in the on state until the device burns out or the power supply is turned off.

Even if latch-up has been extensively studied, it can still represent a problem in the CMOS technologies used in the automotive environment because

1) supply line transients and electrical noise can give rise to parasitic currents that can turn on the parasitic transistors, thus triggering latch-up;
2) latch-up hardness is reduced as the temperature increases;
3) mixed bipolar CMOS technologies may be more sensitive to this phenomenon; and
4) the design of smart-power integrated circuits, which couple high-density CMOS logic with power devices can be quite challenging to ensure immunity from latch-up due to the large voltage swings and high chip temperatures.[8]

To avoid the latch-up problem, we usually implement and electrically characterize special "four-stripe" test structures. These structures mimic the typical layout configurations present in the VLSI (very large scale integration) CMOS technology to be characterized. We can evaluate latch-up sensitivity by measuring the value of the "triggering" current, which has to be injected into I/O lines to turn on the phenomenon.

We can increase latch-up hardness by adopting guard rings, which lower the resistances of substrate and well, shunting the base-emitter junctions of parasitic bipolar transistors (Figure 7b,c). Substrate resistance can be also lowered using a P/P+ epitaxial substrate. Triggering currents higher than 250 mA have been obtained on epitaxial structures with N+ guard rings.

In a finished CMOS IC in which millions of parasitic elements are present, identifying the latch-up site responsible

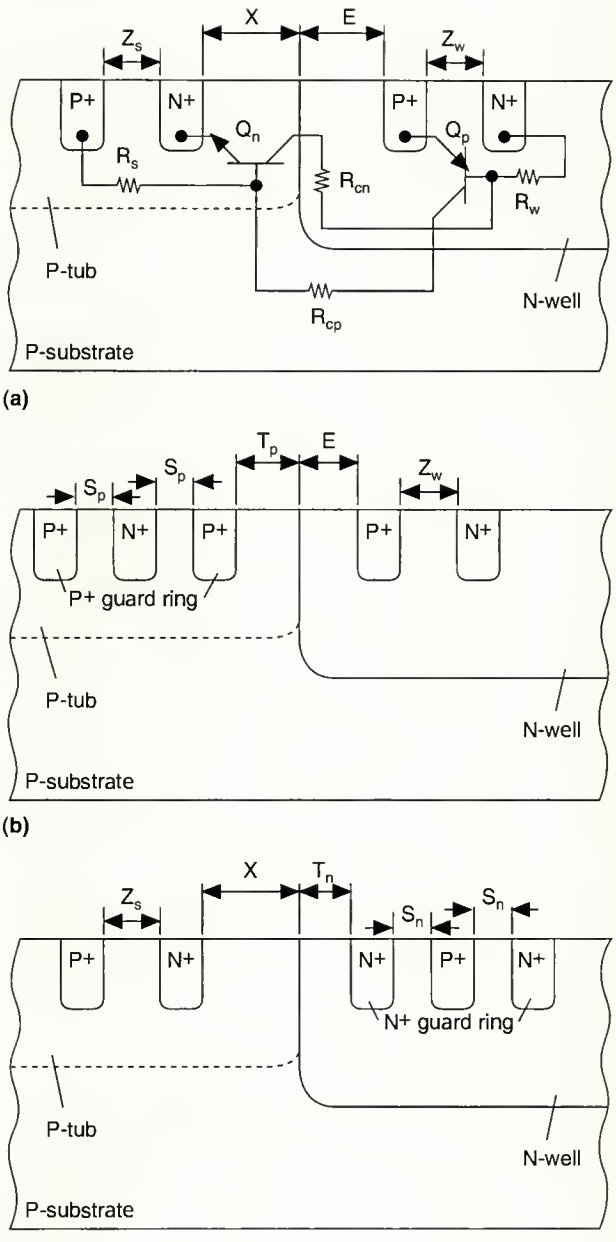


Figure 7. Schematic cross section of latch-up test structure: without guard rings (a), with P+ guard rings in the substrate (b), and with N+ guard rings in the N-well (c).

for circuit malfunctioning can still be easy, if a suitable microscopic technique is adopted. Emission microscopy can detect the infrared light emitted by the forward-biased parasitic transistors due to electron-hole recombination and directly

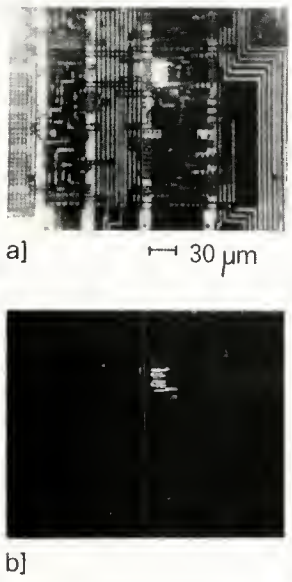


Figure 8. Emission micrography image of a CMOS device in latch-up condition: with topography superimposed (a) and with infrared emission only (b).

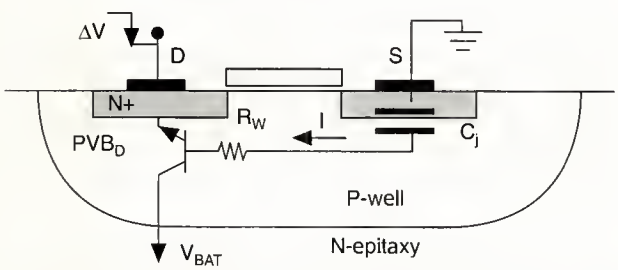


Figure 9. Floating-well concept to avoid latch-up.[11]

point out the failure site.[9] The active parasitic transistors will appear bright in an emission micrograph of the device, biased in the latch-up condition. Figure 8 shows an example in a CMOS EEPROM chip. In Figure 8a the optical micrograph of the device is suprimposed on the infrared emission image; Figure 8b shows infrared emission only. CMOS circuits free of latch-up can be achieved by decoupling the parasitic bipolar transistors using SOI technologies. LAAS CNRS has developed a design methodology based on a floating-well CMOS configuration that prevents latch-up in direct current and transient conditions for a CMOS-compatible smart-power technology, see Figure 9. Providing the well with a two-capacitor, dynamic biasing circuit, completely avoids latch-up initiation due to power device switching or power supply transients.[10,11]

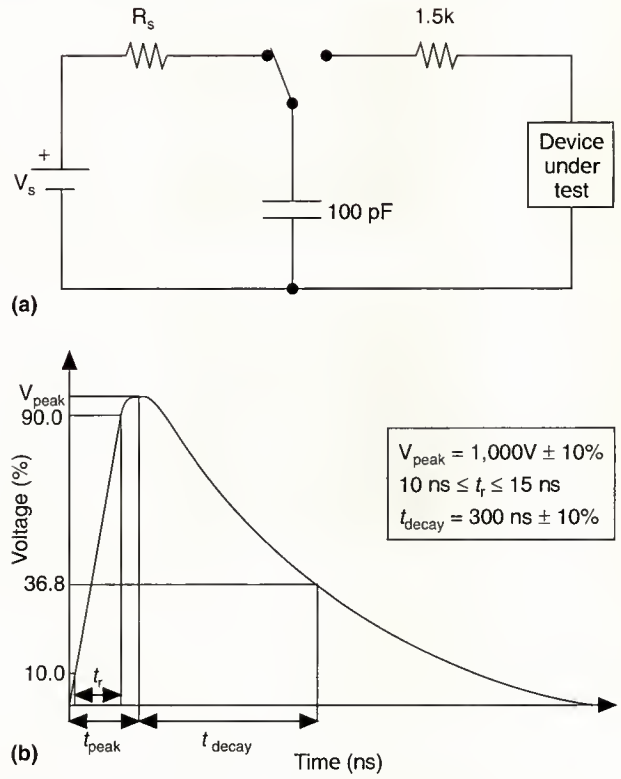


Figure 10. Equivalent circuit of test equipment adopted to simulate ESD stress according to Human Body Model (a) and ESD waveform (b).

Failures due to electrostatic discharge. ESD phenomena produce a reliability concern, which requires careful design of I/O structures.[12,13] Advanced process and device structures may show enhanced ESD sensitivity due to the reduced dimensions, decreased junction depth, and increased breakdown voltages with consequent larger power dissipation during transients. Moreover, maintenance of the electronic systems in the car cannot always be performed while taking all precautions to avoid the risk of ESD.

One possible mechanism of ESD involves a charged body (person working on the line) that discharges through a conductive path into the device (at ground). This most common and completely specified model is known as the Human Body Model; Figure 10a shows its equivalent circuit. The circuit consists of a 100-pF capacitor, which discharges through a 1,500-ohm resistor into the device under test; Figure 10b shows the corresponding waveform.

The research group at Tecnopolis-CSATA together with SGS-Thomson and the University of Padua has studied ESD effects in high-voltage (V_{DS} up to 100V) NMOS and PMOS transistors compatible with CMOS architectures.[13] They developed and tested the following three structures:

- PMOS dual-gate transistors with a P-well drain extension structure; Figure 11 depicts a schematic cross section.
- MOS dual-gate transistors with an N-well drain extension. The device section is the same as in Figure 11, with reverse doping.
- MOS transistors implemented within this same process, using a P insulation implant as the lightly doped drain

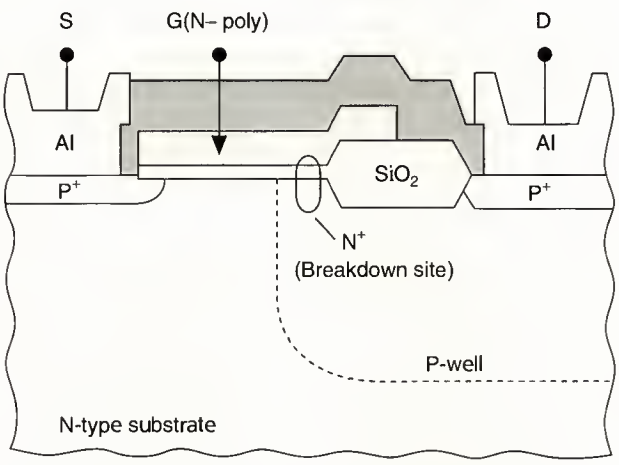


Figure 11. Schematic cross section of the PMOS transistor submitted to ESD testing with the N/P junction induced by ESD between gate and drain.

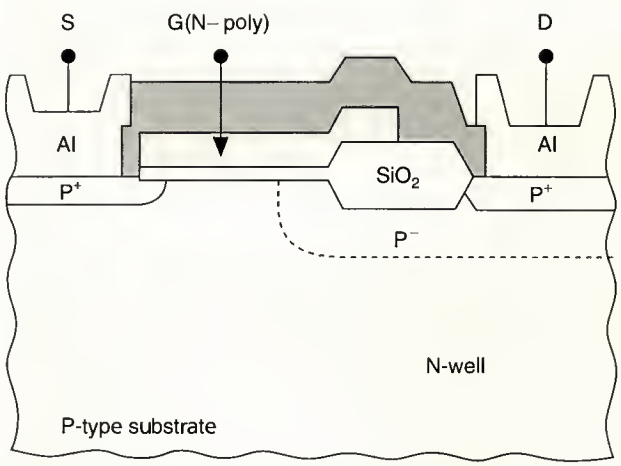


Figure 12. Schematic cross section of the PMOS implemented using the standard P-channel stopper as lightly doped drain regions.

region, that is, adopting the standard P-channel stopper as the drain extension; see Figure 12.

Output transistors were submitted to ESD according to the Human Body Model. They consisted of positive and negative pulses applied to the drain, with source and bulk connected to ground, and the gate grounded through a 1-Mohm resistance. Observed failure modes are

- for PMOS devices, a rectifying junction created between gate (N) and drain (P), see Figure 11, and a parasitic bipolar transistor between gate (N), drain (P), and substrate (N); and
- for NMOS devices, a resistive shunt between gate and source, and a rectifying contact between gate and drain. Threshold voltages for ESD failure are in the 2,500-3,000V range.

To analyze failure, researchers adopted a special Optical Beam Induced Current (OBIC) technique implemented in a laser scanning microscope. Figure 13 is a sketch of the experimental setup employed for OBIC analysis. A scanning laser beam generates electron-hole pairs within the semiconductor. The electric field originated by P-N junctions separates pairs, giving rise to the OBIC current, which is used as the brightness signal on a CRT.

By collecting the signal between the gate and the other electrodes, the sites where a junction has been created or can be accessed due to the failure will appear either bright or dark according to junction polarity. This lets us identify the failure site. Figure 14a shows the OBIC image of a failed PMOS transistor, obtained by connecting the amplifier between drain and bulk. The OBIC signal is collected evenly across the junction, as no defect is present in this area. On the contrary, when the OBIC amplifier is connected between gate and drain (Figure 14b), the signal can be collected only where a P-N junction has been formed, due to ESD, similar to correspondence of the oxide breakdown site between gate and P-well, schematically shown in Figures 11 and 13.

The same research group has studied ESD protection networks suitable for DMOS power transistors[12] and based on lateral NPN transistors or zener diodes (Figure 15, next page). They found that lateral NPN transistors failed at ESD voltages between 2,400V and 3,200V and took emission microscopy images[9] of the failed devices after each step stress.

Figure 16a, next page, shows the emission micrograph of an unstressed NPN lateral transistor biased with a reverse current of 5 μA (in breakdown condition). As can be seen, the emission is evenly distributed and corresponds to the NPN collector-base junction. If a similar micrograph is taken at the same reverse current in the device after the 2,400V ESD stress (Figure 16b), we can see an emission spot that corresponds to the failure site. The dynamic behavior of the tested structure was studied by applying a repetitive, nondestructive square

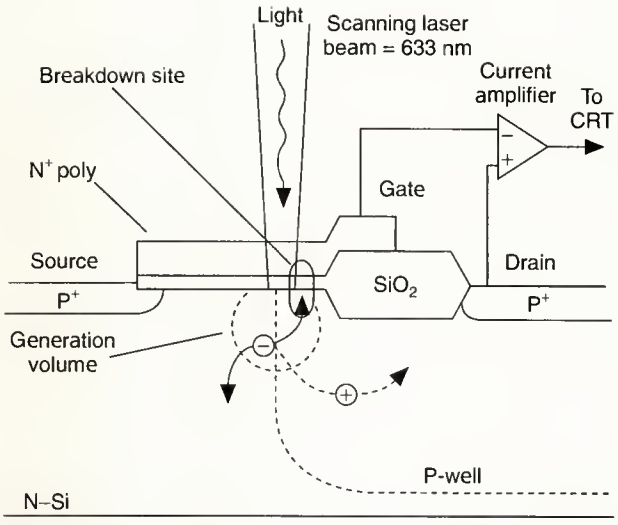


Figure 13. Sketch of the experimental apparatus employed for OBIC analysis of a failed PMOS device.

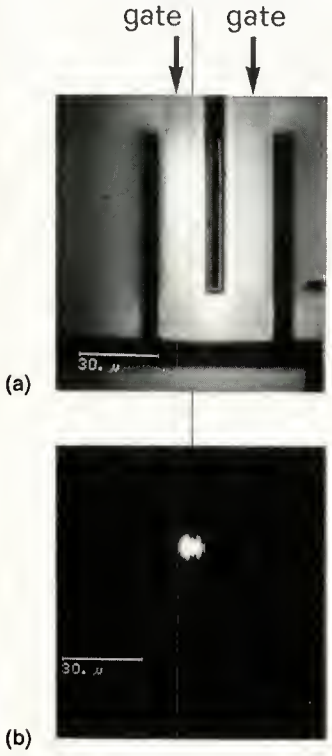


Figure 14. OBIC image of the failed output transistor with OBIC amplifier connected between drain and bulk (a) and gate and drain (b). The white spots correspond to the gate oxide defect induced by ESD.

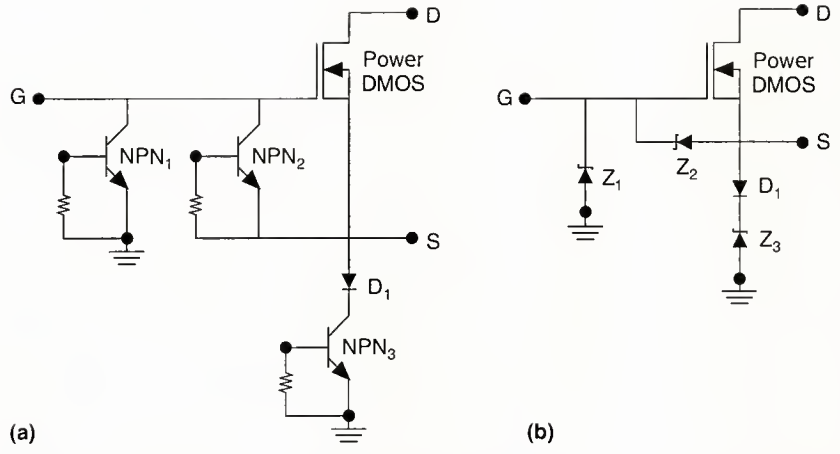


Figure 15. ESD protection networks studied for the DMOS transistor: lateral NPN transistors (a) and zener diodes (b).

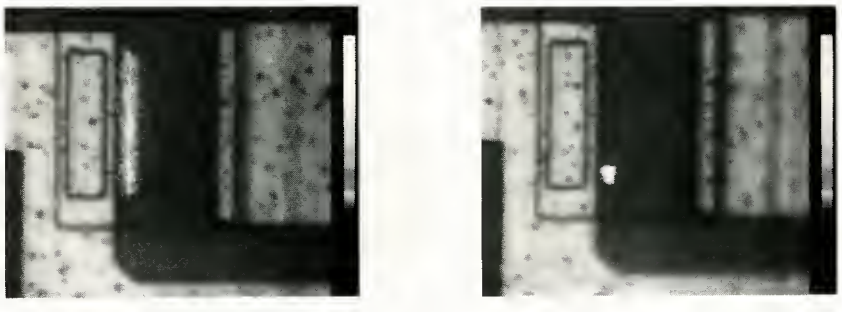


Figure 16. Emission micrograph of a lateral NPN transistor biased with a reverse current of 5 μA in breakdown conditions (a) and after 2,400V ESD stress (b). The spot corresponds to the junction failure site.

voltage pulse to the transistor. The dynamic emission image taken in these conditions, Figure 17, demonstrates that during the pulse most of the current is focused at the BE junction corners, due to the local enhancement of the electric field. This explains the failure location observed in Figure 16b.

The zener protection structure shows a better ESD hardness than the NPN one; in fact, failures are induced only for voltages higher than 4,000V. Moreover, when zener diodes are connected in the network, researchers did not observe failures (either in the protection diodes or in the DMOS transistor) up to 5,000V in the ESD test.

Another ESD model assumes that, during the ESD event, the electrical charge previously accumulated on the device is discharged to ground, thus damaging the device. Grube, Dudek, and Braun at IMS Stuttgart have designed more than 50 I/O protection structures against ESD and load-dump transients, and have tested them according to the described "charged-device" model. Optimized structures having increased ESD hardness have been identified.

The group of Flohrs and Michel at Robert Bosch GmbH has designed a voltage-protected supply input of a smart-ignition coil driver for automotive applications. The power stage switches itself off at voltages exceeding 30V, and it is protected against positive and negative transients on the supply line of an automobile. This research group is currently working on the design of smart-power switches that include diagnostic functions and enable easier fault detection and increased safety against failures.

IC susceptibility to electromagnetic interference. The susceptibility of automotive electronics to EMI can represent a serious threat to the correct operation of electronic systems. Within the car, electronic systems coexist with electrical devices (such as switches, relays, motors, and actuators) that can produce various kinds of electrical noise. This noise can be conveyed on supply and signal lines, forcing electronic circuits to operate incorrectly. Moreover, lightning events, radio and TV transmitters, and radar systems are sources of intense electromagnetic radiation; we can encounter one of these sources when driving close to an airport or a long-distance broadcasting station.

If the RF signals are extremely intense, electronic devices can even fail catastrophically due to the induced temperature rise. In this case, several failure mechanisms such as metal-semiconductor interdiffusion and short-circuiting of shallow junctions in ICs can be induced. Less intense signals can bring about temporary circuit malfunctioning. Since the experimental characterization of these effects on the electronic systems mounted in a car is extremely difficult, researchers are trying to develop modeling and testing techniques to evaluate EMI effects on relatively simple circuit elements. The results can be used to improve system and device design to reduce susceptibility to EMI.

The problem can be divided into two tasks. First we evalu-

ate the coupling between the incoming radiation and the car electronic system to determine how much interfering power is conducted into the terminals of the ICs employed. Then we need to evaluate the susceptibility of each IC to conducted radiation, that is, to signals directly applied to device I/O and supply lines. The first task requires extended experimental characterization of the different sources of electromagnetic noise to which the car can be subjected. The second problem has been analyzed by several researchers, starting from the publication of the *IC Susceptibility Handbook* developed by the US National Aeronautics and Space Administration and McDonnell Douglas in 1978.[14]

Pozzolo and coworkers at the Politecnico di Torino in Italy studied the susceptibility of ICs for automotive applications to EMI. They aimed to develop design tools that take into account EMI problems during the development phase of new electronic products. They studied the susceptibility of active devices to EMI by

- making measurements on different devices to find the power level of the interference signal required to have susceptibility at the different frequencies; and
- defining suitable models for the active devices, which enable a simulation of the device behavior in the presence of an interference signal to be performed.

Simulating the effect of electromagnetic interference on ICs gives us information about the more important parameters so we can design devices with high immunity.[15,16]

To study the susceptibility problem at printed board and device levels, researchers have developed models both for bipolar junction transistors and for field-effect devices. These models describe the behavior of devices subjected to conducted EMI at device terminals; in this way, a single linear simulation substitutes several nonlinear analyses in the time domain. A complete macro model for the study of the susceptibility of operational amplifiers to EMI has been developed.

The authors also characterized the filtering action of different packages and mounting techniques by exploiting the time domain capabilities of a network analyzer. The technique has proved to be very useful in separating the influence of printed board interconnections from the circuit model of the package interconnections and bonding. Several experiments were carried out on operation amplifiers with different packages, confirming the validity of the approach.

Developing high- and medium-voltage MOS technologies. Several applications of automotive electronics, ranging from display drivers to intelligent power actuators for multiplex wiring systems, require the development of reliable power MOS devices. These MOS devices must withstand drain voltages in the 10-200V region. Several groups within PRO-CHIP have therefore studied the performance and the reliability of high- and medium-voltage MOS transistors.

Figure 17. Dynamic emission micrograph taken when a positive square voltage is applied to the tested lateral NPN transistor.

Different approaches have been followed. Bafleur and coworkers at LAAS-CNRS have developed an N-channel vertical DMOS technology on an N epitaxial layer whose thickness is related to the device's voltage-handling capability (10 µm for 60V), compatible with a CMOS process. This technology adopts a floating-well concept with capacitance coupling to reduce latch-up, and a low-doped drain technology for the low-voltage NMOS and PMOS transistors. This technology also reduces the electric field in the channel region, thus limiting hot-electron effects and improving breakdown voltages.[17,18] The group is currently working toward the integration of an electrical motor control circuit for automotive applications in BiCMOS technology.

Ifström and coworkers at IMS Stuttgart used thermal bonding of oxidized wafers to obtain a high-quality SOI substrate, useful both for electronics and sensor applications. With this substrate, self-isolated lateral and vertical DMOS transistors can be achieved (Figure 18).[19,20] A smart-power process containing 150V, 0.8-µm VDMOS, 2-µm CMOS and bipolar devices with full dielectric isolation on fusion-bonded SOI has been developed. Vertical power devices can be obtained by silicon direct bonding of Si_3N_4 to SiO_2, exploiting the bonded nitride layer as a selective etch stop. Despite not passivating the surface, a breakdown voltage of over 500V was obtained.

SOI technologies improve device reliability in different ways:

- dielectric isolation enables latch-up to be avoided completely, at least in buffer stages;

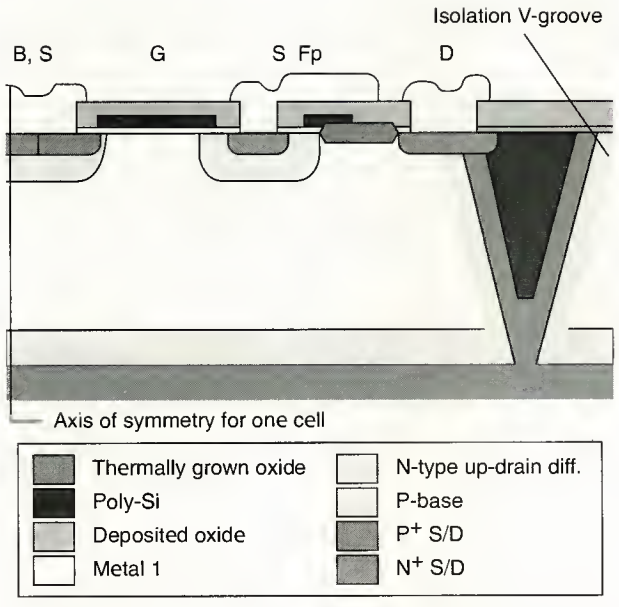


Figure 18. Cross section of a VDMOS transistor implemented in thin SOI with the direct bonding technique.

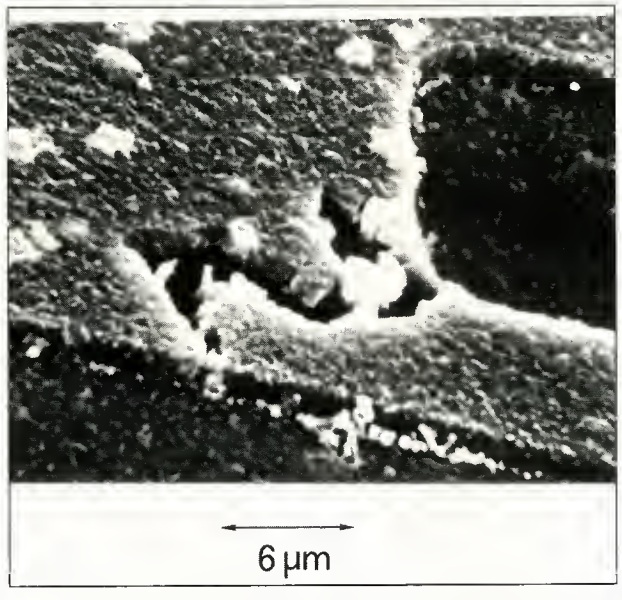


Figure 19. Void in the AlCuSi metallization of an emitter-coupled-logic IC induced by electromigration.

- the leakage current is reduced, making low-level operation at elevated temperatures possible; and
- the management of parasitics is simplified.

Electromigration. Electromigration can be a significant cause of failure in metallic films used as interconnections in electronic circuits. Due to the interaction with flowing electrons, the atoms of the IC metallization tend to migrate toward the positive end of the conductor. If this material transport is not homogeneous, creation of voids and material pile up can occur, resulting in open circuits, as shown in Figure 19, or in short-circuiting between overimposed metal layers. The continuing trend toward scaling down device dimensions has led to a drastic reduction in the metal line sections and in contact areas, increasing the risk of electromigration, which is accelerated by high current densities.

The OBIC technique previously described can also be successfully applied to the study of supply metal interruptions due to corrosion or electromigration in complex ICs. In this case the electron-hole pairs generated by the scanning laser beam are separated by the drain (source) junctions of MOS transistors. The V_{DD} and V_{SS} contacts collect the generated carriers. A current can therefore be detected by the OBIC amplifier connected at the supply terminals, thus generating a contrast in the OBIC image. Those device regions with supply line interruptions do not contribute to the OBIC signal.

Researchers can easily identify the failure sites by comparing a failed and a functional device. An example is shown in Figure 20; the OBIC image on top refers to the unfailed device, while the micrograph below refers to a failed one. Due to an interruption in the V_{SS} metal (black circle), all devices connected to the corresponding branch of the supply line are not biased and do not give rise to black contrast in the OBIC image. Because the interrupted V_{SS} line only supplies the device internal RAM, automatic testing detected only a functional failure. Supply line interruptions were always found on oxide steps or where current density suddenly increases, due to the decrease in metal width, or to the presence of corners. This finding strongly suggests that electromigration has taken place in these devices. The technique enables failure sites to be quickly identified, thus cutting the costs and the time required for failure analysis of complex circuits.

Reliability of GaAs devices. Some interest in III-V devices for automotive applications has arisen recently for three main reasons:

- carrier frequencies around 60 GHz are envisaged in Europe for road-to-car and car-to-road transmission;
- collision avoidance radars will most probably use 76-77 GHz; and
- GaAs ICs can be operated at higher temperatures than silicon, due to the larger energy gap of GaAs with re-

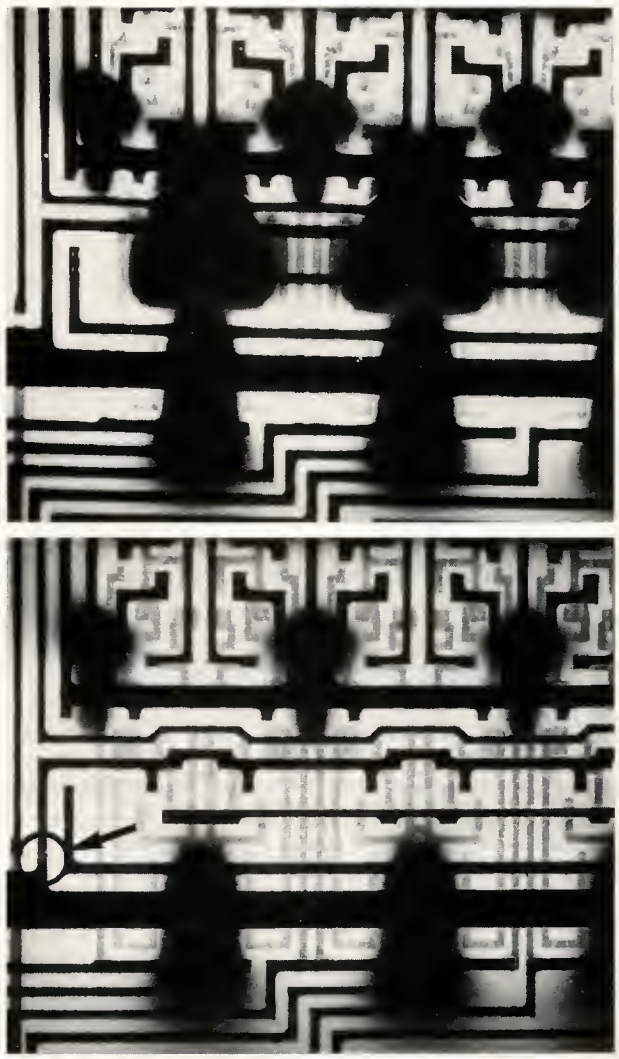


Figure 20. OBIC image of a portion of a microprocessor: unfailed samples (top) and failed sample returned from the field (bottom). The OBIC contrast reveals that the supply line is interrupted (possibly due to electromigration) in the area indicated by the black circle.

spect to silicon, even if this last advantage is partly compensated by a lower substrate thermal conductivity.

Microwave applications would possibly lead in the future to the use of low-noise and power GaAs MESFETs and high-electron mobility (heterojunction) transistors (HEMT) in the automotive environment.

The schematic cross section in Figure 21 shows the typical structure of an AlGaAs/GaAs HEMT device. In the

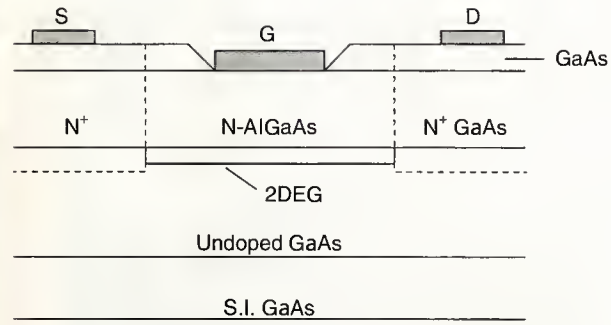


Figure 21. Schematic cross section of an AlGaAs/GaAs high electron mobility transistor (S.I. = semi-insulating.)

heterostructure, the AlGaAs "donor" layer is N doped, while the GaAs layer is not intentionally doped. Electrons transfer from the wider energy-gap material (AlGaAs) into GaAs near the heterostructure interface, forming a two-dimensional electron gas (2DEG). In this way, carriers are separated from ionized impurities, thus avoiding impurity scattering and achieving higher carrier mobilities.

This fact, together with the reduced distance between the conducting channel and the gate electrode, leads to higher values of transconductance, higher operating frequencies, and better noise characteristics in HEMTs compared to MESFETs. For this reason, HEMTs are replacing conventional low-noise MESFETs in MMIC (monolithic microwave IC) technologies, and the study of their long-term stability has become relevant. In addition, critical issues are related with the stability of multilayer metallizations used for Schottky gates and ohmic contacts, the dopant redistribution in the semiconductor, and the presence of electron traps (deep levels) in the AlGaAs layer and of surface states.

The University of Padua in cooperation with Alcatel Telettra SpA has investigated the reliability of commercially available AlGaAs/GaAs HEMTs from four different suppliers by means of a storage at T = 225-275°C and of biased life tests.[21] The main technological differences among the devices concern the gate metallization. Two device types (A, E) have Al/Ti gates, type B has Al/Ni gates, while supplier C adopted a gate metallization based on refractory metal silicide (Au/Pt/Ti/WSi). Figure 22 (next page) shows the cross-section transmission electron micrograph of a HEMT device with Al/Ti gate metallization.

The main failure mechanisms detected are the

- increase of Schottky barrier height of the gate diode Φ_B in devices (type B) with an Al/Ni gate;
- decrease of Φ_B in devices (type A and type E) with an Al/Ti gate; and

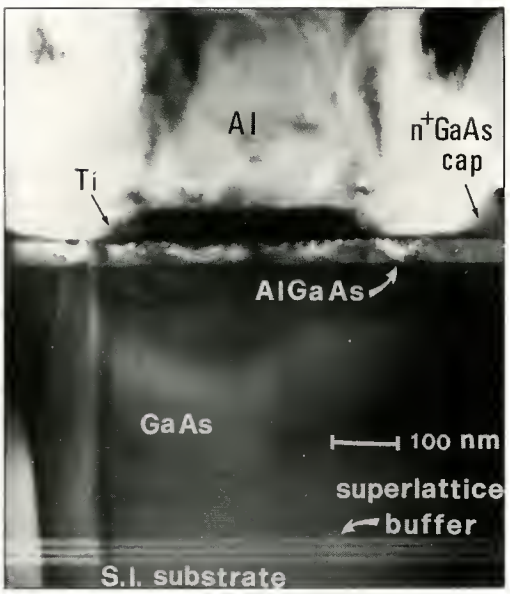


Figure 22. Cross-section transmission electron micrograph of a HEMT device with Al/Ti gate metallization.

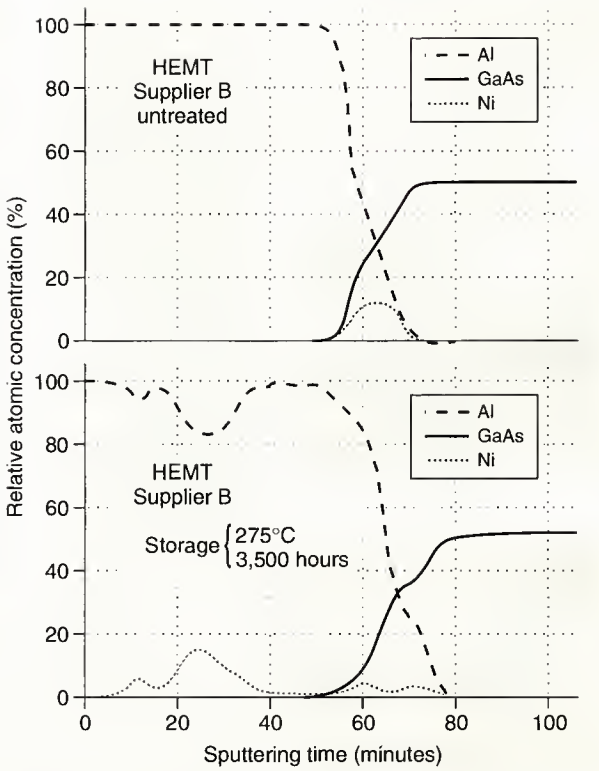


Figure 23. Auger atomic in-depth profiles of Al/Ni gate metallization in an untreated HEMT sample (top) and in one sample aged for 3,500 hours at 275°C (bottom).

- increase of parasitic resistances of source/drain contacts in type A and type E devices.

These failure mechanisms are thermally activated, and degradation rates have been found to be linearly proportional to the square root of annealing time. Failure times t_f were found to follow an Arrhenius dependence on temperature, $t_f = A \exp(-E_a/kT)$, where k is the Boltzmann constant, A is a constant, T is the absolute temperature, and E_a is the activation energy, characteristic of the degradation mechanism considered. Al/Ni gate contacts presented an increase of barrier height with E_a = 2.0 eV, while Al/Ti gate contacts show a decrease of barrier height with E_a = 1.3 eV. An increase of source and drain parasitic resistances has been detected in devices of two suppliers with E_a = 1.6 eV.

To identify the physical reasons for the observed changes in Φ_B, researchers adopted Auger electron spectroscopy. This technique can follow the in-depth atomic profile of the various elements that form the metal/semiconductor contact, thus detecting possible interdiffusion effects. Analyses have been performed on untreated and aged devices.

Figure 23 shows results obtained after an aging period of 3,500 hours at 275°C on Al/Ni devices as a significant example. The as-received devices showed a thin Ni film concentrated near the interface between the Al metallization and the semiconductor substrate. The Auger in-depth profile reported in Figure 23 indicates that Ni has been evenly redistributed through the Al film during the aging test, reaching a concentration value which is barely over the detection limit of the technique (1 percent). Ni likely forms a saturated solid solution in Al. Interfacial reactions between Al and GaAs are also clearly detected, with a long Al diffusion tail into the semiconductor substrate. Reactions at the Al/GaAs interface are well-known to induce an increase of the gate diode barrier height, as observed in this case. Despite the presence of these failure mechanisms, comparison with tests on low-noise MESFETs does not show major reliability problems for heterostructure devices.

Reliability prediction and reliability data banks. Even if the approach of "measuring reliability" becomes obsolete as the failure rate of electronic components decreases, designers still need reference data for estimation of system reliability, calculations of cost of spare parts and of repairs, evaluation of warranty periods, and comparison of different designs. For military electronic equipment and systems, *MIL Handbook 217* is the standard for reliability predictions; however, its applicability to other environments is often discussed. In fact, even if based on a large amount of reliability data, predictions of the MIL handbook are often too conservative, leading to overesti-

mation of failure rates and to costly overdesign. New reliability models are being proposed for the new version of the handbook, and the importance of a more detailed knowledge of the physics of failure mechanisms of electronic components is being stressed in a new proposal for the standardization of reliability testing and device screening procedures.[22]

In the United States, the Society of Automotive Engineers has proposed prediction techniques for automotive applications, which can be used as a reference for general reliability evaluation of electronic components.[23] A constant failure rate λ_p is assumed, which is calculated on the basis of a mission profile of 400 hours/year, and is directly calculated from the formula

$$\lambda_p = \lambda_b \Pi_F \Pi_S \Pi_P \Pi_T$$

where λ_b is the basic failure rate, Π_F is a factor identified by the component type, Π_S is the screening factor, Π_P is a factor identified by the package type, and Π_T contains the temperature dependence of the failure rate. The calculation of the factor Π_T is based on the Arrhenius law and uses activation energies that depend uniquely on the device technology (for example, 0.4eV for all digital bipolar ICs, 0.7 for all MOS logic circuits, and so on).

In Italy, the PRO-CHIP research unit at Tecnopolis is responsible for the operation of the Reliability Circle, a nonprofit organization. It is joined by the main electronic component and system makers and users, and focuses its activity on the exchange of data and experience concerning the reliability of electronic components and the related techniques. The Reliability Data Bank contains more than 3,000 reports. The Circle promotes special meetings for the exchange of information and the definition of common methodologies for quality assurance, testing, and reliability.

The researchers designed a data bank in the PRO-CHIP subproject for the collection of reliability data concerning automotive electronics. The data bank, based on SAE-defined models, allows reliability predictions according to SAE and MIL-STD models. The analysis of multisource reliability data has been performed by means of classical and Bayesian statistical approaches. It confirmed the electronic component reliability trend, especially that concerning field failure data, to be congruent with the estimates calculated by the *MIL Handbook 217F* model in standard conditions.[24]

Fail-safe operation

Safety-critical automotive electronics tasks such as steering and braking control and collision avoidance require fail-safe or fault-tolerant components. Fail-safe operation of a system avoids the dangerous consequences of a fault by switching into a "safe" state; in other words, a fail-safe system either works correctly or is in a safe condition. A fault-tolerant system works correctly even in the presence of a fault; that is, it detects the fault and corrects the related errors. The extensive use of fail-safe or fault-tolerant techniques is not possible at this moment within automotive systems, since it would introduce excessive overhead and costs. Such use will become mandatory in the near future for critical parts and subsystems, requiring specific design techniques. Table 2 (next page) summarizes the current PRO-CHIP research in this area.

Characterization of metastable behavior of bistable devices. Marginal triggering conditions can place bistable devices in metastable conditions. Two types of metastability can occur: analog and oscillatory. The former causes the output of the device to stay at an electrical level near the input threshold voltage, while the latter causes the output to toggle repeatedly between the two opposite logic levels. Metastability is unavoidable, but its effects can be evaluated and limited within known bounds by using appropriate design methods. A complete understanding of the metastability is therefore an essential step in the design of devices that are intrinsically safer. Del Corso and coworkers[25] have studied oscillatory metastability, developing analytical models. These models let us understand circuit parameters and electrical conditions that trigger metastable oscillations so we can identify them and improve the resolving time of oscillations.

Electrical and optical CAN. A CMOS driver has been developed for the automotive controller area network (a protocol, implemented in hardware only and specially designed for automotive applications).[26] In addition to a single external component the device can withstand 120V load-dump transients, 0.33A-24V shorts, and latch-up triggering currents up to 1A, 0.1s.

An all-optical network has also been implemented, which offers very high immunity to electromagnetic interference. The adopted ring topology enables various failures to be identified; the network can tolerate faults by means of a redundant structure, coupled to supervising circuits.

Fail-safe processor. IMS[27] designed a fail-safe VLSI controller, minimizing area requirements by using optimized combinations of duplicated units and error coding. A structured approach lets users analyze possible hardware faults on a high level; stuck-at and bridging faults have been considered. A duplicate ALU in the controller avoids complex error coding, while RAM and ROM are implemented with error-detection mechanisms (Figure 24, next page). The processor consists of 20,000 transistors and has a peak performance of 10 MIPS at a maximum frequency of 20 MHz. Plans call for the next processor version to include on-line error detection by means of the instruction sequence check method.

IN DESCRIBING THE RELIABILITY RESEARCH ACTIVITIES within the PRO-CHIP project, we mentioned investigations of both new reliability assessment methodologies and intrin-

Table 2. Fail-safe operation research groups within Prometheus PRO-CHIP.

Research project	Institution	In cooperation	Device/system studied	How fail-safe operation is achieved
Hardware for communication interfaces	Steinbeis TZ Mikroelektronik und Systemtechnik-Furtwangen, Germany	Robert Bosch GmbH	Electrical and optical controller area network	Rugged ACMOS technology; optical network with ring topology
Design of IC for concurrent error detection	Dip. di Ing. Elettronica, 2nd Univ. of Rome, Tor Vergata, Italy	—	Electronic subsystems	Self-checking circuits for logic; residue theory for arithmetic, correcting codes for memory
Fail-safe systems	Institute for Microelectronics Stuttgart, Germany	Daimler Benz AG	VLSI controllers Electr. steering demonstrator	ALU redundancy; error correction in RAM/ROM; triple redundancy with vote
Characterization of metastable behavior of bistable devices	Politecnico di Torino, Italy	—	Bistable devices	—

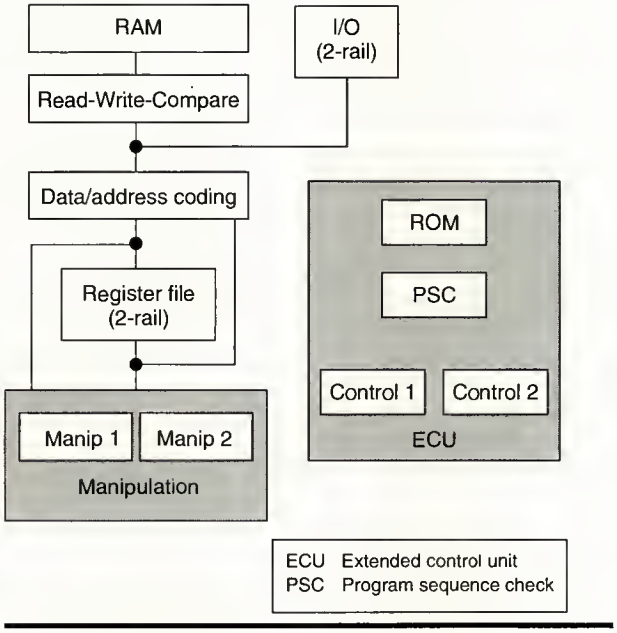


Figure 24. Minimized fail-safe system.[27]

sically reliable device technologies and designs. In particular, new I/O protection networks, smart-power devices, and fault-tolerant and fail-safe architectures are being developed, to reach the reliability requirements imposed by complex electronic systems to be used in future cars.

Other areas must be investigated to further improve safety, such as failure mechanisms and reliability of sensors and actuators; assessment of software reliability is necessary for critical applications, such as collision avoidance, automatic steering, and braking control. Rugged smart-power technologies have already found many applications within the car, and we can envisage that the use of fault-tolerant and fail-safe controllers for automotive applications will become increasingly popular in the next decade.

Acknowledgments

We thank all colleagues in the PRO-CHIP community who have provided their results and useful comments and suggestions. Particular thanks go to Hans Herrmann and Bernd Hoefflinger (both at IMS Stuttgart, Germany), and Marise Bafleur (CNRS LAAS, Toulouse, France). We also thank Marianna Cavone, Roberto Rivoir, Michele Stucchi (Tecnopolis CSATA, Bari, Italy), Letizia Bertolini, and Emanuela Zoccolanti

(Marelli Autronica, Pavia, Italy) for providing their data and experience on failure analysis techniques. Finally, a special acknowledgment goes to Marie English (IEEE Computer Society) who carefully read the manuscript and gave many suggestions to improve the readability and effectiveness of this article. The responsibility for any weakness in this article remains with the authors, who apologize for possible errors in describing PRO-CHIP research activities of other colleagues.

CNR, Progetto Finalizzato Trasporti (a project of the Italian National Council of Research Finalized to the improvement of means of Transportation) and the Eureka project Prometheus partially supported this work.

References

1. B. Hofflinger, "The Eureka Program for European Traffic with Highest Efficiency and Unprecedented Safety (Prometheus) and Some of Its Reliability Issues," *Proc. Third European Symp. Reliability of Electron Devices, Failure Physics and Analysis (ESREF)*, vde-verlag, GmbH, Berlin und Offenbach Bismarckstrasse 33, D-1000 Berlin 12, Germany, 1992, pp. 108-110.
2. H. Herrmann, "Reliability Research in the PRO-CHIP Program of Prometheus," *Proc. ESREF,* 1992, pp. 111-131.
3. S.J. Flint and L.L. Kent, "Electronic Engine Control: Automakers Contend with One of the Harshest Application Environments," *IEEE Spectrum,* 1981, pp. 61-62.
4. J.E. Semmens and L.W. Kessler, "Nondestructive Evaluation of Thermally Shocked Plastic IC Packages Using Acoustic Microscopy," *Proc. 29th Ann. IEEE Int'l Reliabilty Symp., ISTFA*, 1991, pp. 160-166.
5. A. Hijazi et al., "Thermomechanical Failure Analysis of SMD Solder Joints," *Proc. ESREF,* 1992, pp. 327-330.
6. A. Napieralski et al., "Developing a True 3D Transient Thermal Model for Power Components or Hybrid Power Circuits," *Proc. Fourth European Conf. Power Electronics and Applications, jointly with the Symp. Materials and Devices for Power Electronics (EPE-MADEP),* 1991, pp. 104–109.
7. P. Leturcq el al., "A New Approach to Thermal Analysis of Power Devices," *IEEE Trans. Electron Devices,* 1987, pp. 1147-1156.
8. C. Canali et al., "Infrared Microscopy Study of Anomalous Latch-up Characteristics Due to Current Redistribution in Different Parasitic Paths," *IEEE Trans. Electron Devices,* 1989, pp. 969-978.
9. J. Kolzer et al., "Quantitative Emission Microscopy," *J. Applied Physics*, 1992, pp. R23-R41.
10. M. Bafleur et al., "Capacitive Coupling of Floating Body Well to Sensitive Nodes Prevents High-Resistance CMOS Circuit from Latching-up," *Electronic Letters,* 1989, pp. 860-861.
11. M.P. Vidal et al., "Floating Well-Based Design Methodology Aimed to Improve Latch-up Immunity in a Smart Power Technology," *Microelectronic Engineering,* 1992, pp. 161-164.
12. B. Bonati et al., "Analysis of ESD Protection Networks for DMOS Power Transistors by Means of Static and Time-Resolved Emission Microscopy," *Proc. ESREF,* 1992, pp. 73-76.
13. E. Zanoni et al., "Detection and Localization of Gate Oxide Shorts in MOS Transistors by Optical Beam Induced Current," *IEEE Trans. Electron Devices,* 1991, pp. 417-419.
14. McDonnell Douglas, *Integrated Circuit Susceptibility Handbook*, 1978.
15. V. Pozzolo, "Susceptibility of Integrated Circuits to Electromagnetic Interference," *Proc. Third Prometheus PRO-CHIP Workshop,* May 1990, pp. 15-38.
16. V. Pozzolo et al., "Susceptibility of ICs to EMI," *Proc. Third Prometheus PRO-CHIP Workshop,* May 1990, pp. 189-194.
17. A. Elmoznin et al., "The Smart Power High-Side Switch: Description of a Specific Technology, Its Basic Devices, and Monitoring Circuitries," *IEEE Trans. Electon Devices*, 1990, pp. 1154-1161.
18. M. Bafleur et al., "Driving and Protection Circuitry for a Smart Power MOS High-Side Switch Based on a Floating-Well Concept," *Proc. EPE-MADEP,* 1991, pp. 189-192.
19. U. Apel et al., " A 100-V Lateral DMOS Transistor with a 0.3-mm Channel in a 1-mm Silicon-Film-on-Insulator-on-Silicon," *IEEE Trans. Electron Devices,* 1991, pp. 1655-1659.
20. T. Ifström et al., "A 150-V Multiple Up-Drain VDMOS, CMOS, and Bipolar Process in Direct Bonded Silicon on Insulator on Silicon," *IEEE Electron Dev. Lett.,* Vol. EDL-13, No. 9, 1992, pp. 460-461.
21. C. Canali et al., "Reliability Aspects of AlGaAs/GaAs HEMTs," *Proc. 29th Annual IEEE Int'l Reliability Physics Symp.,* 1991, pp. 206-213.
22. G.F. Weston, "MIL Reliability: A New Approach," *IEEE Spectrum,* Aug. 1992, pp. 46-49.
23. SAE, Electronics Reliability Subcommittee, *Automotive Electronics Reliability Handbook*, AE-9, Soc. of Automotive Engineers, Warrendale, Penn., 1987.
24. M. Chirulli et al., "A Combined Statistical Analysis of Multisource Reliability Data of Electronic Components," *Proc. ESREF,* 1992, pp. 215-219.
25. L. M. Reyneri, D. Del Corso, and B. Sacco, "Oscillatory Metastability in Homogeneous and Inhomogeneous Flip-Flops," *IEEE J. Solid-State Circuits,* 1990, pp. 254-264.
26. W. Kuntz and R. Mores, "Hardware for Communication Interfaces," *Proc. Sixth Prometheus PRO-CHIP Workshop,* May 1992, pp. 168-170.
27. M. Jurczyk and T. Schwederski, "Fail-Safe Systems," *Proc. Prometheus PRO-CHIP Workshop,* May 1992, pp. 82-84.

Additional readings

Readers interested in reading further on these topics may contact the author (address at the end of this article) for an extended list.

Enrico Zanoni is an associate professor of Applied Electronics in the Department of Electronics and Computer Science at the University of Padua. His research activities include the study of the reliability of silicon and compound semiconductor devices, improvement of MOS processes for medium and high voltages, and experimetnal characterization and modeling of impact ionization effects in silicon-based and gallium arsenide-based bipolar devices and FETs. He has cooperated with the Eureka project Prometheus since 1987, analyzing reliability aspects of automotive electronics within the PRO-CHIP subproject. Zanoni graduated in physics at the University of Modena. He has authored or coauthored more than 100 papers published in international journals or presented at conferences, and four book chapters. He is a member of AEI (Italian Electrotechnical Association) and the IEEE and the Computer Society.

Paolo Pavan is a PhD student in the Electronics and Computer Science Department of the same university. He is currently at the University of California, Berkeley, within the framework of a student exchange program between the University of California and the University of Padua. Earlier he obtained a research grant from SGS-Thomson to study latch-up effects in CMOS ICs. His research activity includes study of the reliability of CMOS ICs and characterization of impact-ionization effects in silicon, SiGe, and AlGaAs/GaAs bipolar transistors.Pavan graduated in electronics engineering at the University of Padua.

Direct any questions concerning this article to Enrico Zanoni, Dipartimento di Elettronica e Informatica, Universita di Padova, Via Gradenigo 6A, 35131, Padova, Italy; or e-mail at zanoni@elena.dei.unipd.it.

Reader Interest Survey

Indicate your interest in this article by circling the appropriate number on the Reader Service Card.

Low 159 Medium 160 High 161

Vision Assistance in Scenes with Extreme Contrast

Applications of vision systems in traffic environments still suffer from the limited optical dynamic range of their sensors and lack of flexibility in readout mechanisms. We describe the performance and architecture of a High Dynamic Range Camera (HDRC) chip and the conceptional advantages for its adaptation to image processing systems.

Ulrich Seger

Heinz-Gerd Graf

Institute for Microelectronics Stuttgart

Marc E. Landgraf

Intel Corporation

Several applications of image processing systems are under development within the European Prometheus project, which is a cooperative research program.[1] The task of these image processing systems is to deliver actual, well-organized, and highly reliable data to the driver but also to driver assistant systems. The assistant systems help to keep a car in its lane, recognize obstacles, or enhance visibility under certain circumstances.

If image data are to be used in vehicle control or warning systems, they must support short response times. For example, steering processes require response within some milliseconds. Imaging of high-contrast scenes with brightness changes of 100,000:1 from frame to frame is necessary for uninterrupted processing without delays. However, this is not possible with changing apertures or varying shutter or integration times.[2]

Commonly available cameras with an optical dynamic range of about 5,000:1 (74 dB), and even high-performance devices known from the literature[3,4] to reach 8,000:1 (78 dB), fall short of the minimum dynamic range of 100 dB desired in automotive applications. (This dynamic range is necessary to avoid severe saturation, caused by reflections of bright sources such as the sun.)

Some camera system approaches attain a higher dynamic range by controlling shutter, aperture, or signal integration time, but may struggle with oscillations under rapidly changing conditions. (Imagine the effects created by the shadows in a tree-lined road.) These system approaches require extra exposure control and image postprocessing hardware as well as extra time for subsequent readout and image reconstruction.

Help may come from a combination of a hardware-implemented logarithmic signal compression with a RAM-like pixel access and the opportunity to integrate such circuits together with application-specific signal postprocessors into a standard CMOS process. This approach leads to higher system performances in applications in which high scene contrast is a problem.

Sensor architecture

During the development of the HDRC (High Dynamic Range Camera) chip, we placed special emphasis on a processor-friendly architecture. Systems engineers should be able to benefit from high optical performance as well as from an image sensor interface that is easy to adapt. Pixel processors implemented within the focal plane enlarge the application field toward imaging of extreme contrast scenes, and a RAM-like digital interface supports random access to each pixel with a minimum access time of 150 ns. A nondestructive readout mechanism allows subsequent access to the same pixel at even higher frequencies. (Figure 1 shows the HDRC64 sensor architecture, the version with 64 × 64 pixels and our prototype.)

0272-1732/93/0200-0050$03.00 © 1993 IEEE

A maximum readout frequency of 6.6 MHz allows frame rates of above 1,600 frames/second (using the full 64×64-pixel field) but can even reach higher frame rates when accessing a smaller area of interest.

The total data rate may be further increased with a multifield architecture (see Figure 2), which supports multiple parallel outputs and therefore may serve as an input device for processor arrays.

To participate in the further scaling of technology and in design enhancements of digital macrocells, we used a standard CMOS technology as the target technology.

Table 1 lists the specifications of the HDRC64.

Local pixel processor

This processor, which is placed around each pixel (see Figure 3, next page) within the focal plane, performs a logarithmic signal compression directly at the place of signal generation.[5] This arrangement prevents an information loss, which might occur should any of the preceding signal transport or processing circuits become saturated.

A logarithmic compression technique known from most biological systems shows some advantages concerning the dynamic range of input signals that may be processed.

The HDRC chip achieves logarithmic compression by controlled draining of the photocurrent that normally would contribute to an output voltage proportional to the irradiated power. Chamberlain first used this technique in the early 1980s.[6] A development toward higher robustness and compatibility with today's CMOS technologies resulted in a different conversion principle of the pixel processor, but it still converts an input signal to its logarithm at the pixel output. Also, the local pixel processor simplifies the implementation of area arrays by supporting full addressing capabilities to each pixel.

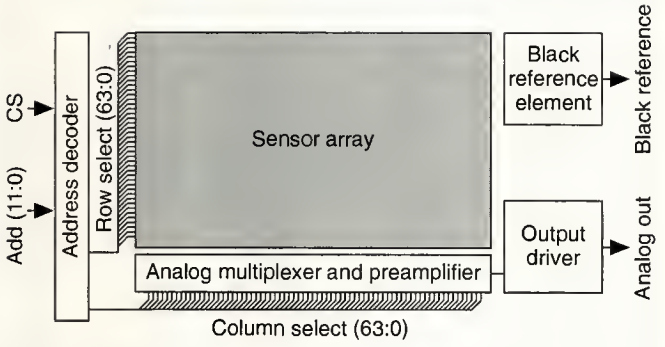


Figure 1. HDRC64 sensor architecture.

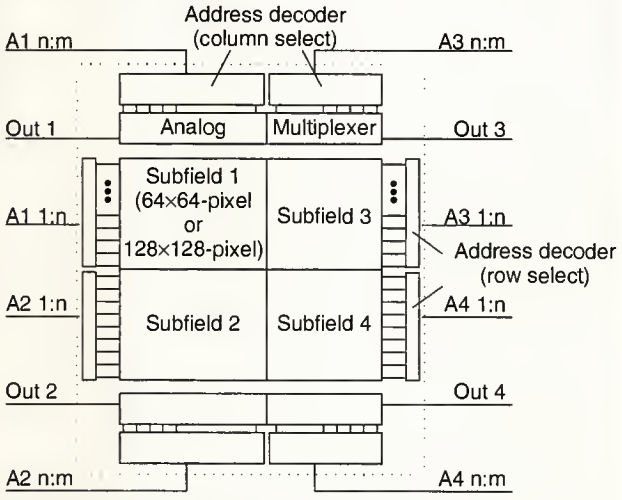


Figure 2. Multifield architecture.

Table 1. HDRC64 specifications.

Parameter	Minimum	Typical	Maximum	Unit
Power supply +	—	5	—	V
Power supply –	—	0	—	V
Quiescent current total chip	—	12	—	mA
Operating current at 1-MHz readout frequency	—	19	—	mA
Pixel count	—	64 × 64	—	—
Total photosensitive area	—	3.84	—	mm^2
Fill factor*	—	> 40	—	%
Optical input signal dynamic range	—	1:100,000	—	—
Resolvable contrast	—	10	—	%
Repetitive pixel readout frequency**	DC	—	6.6	MHz

*In active area ** Depends on incident power

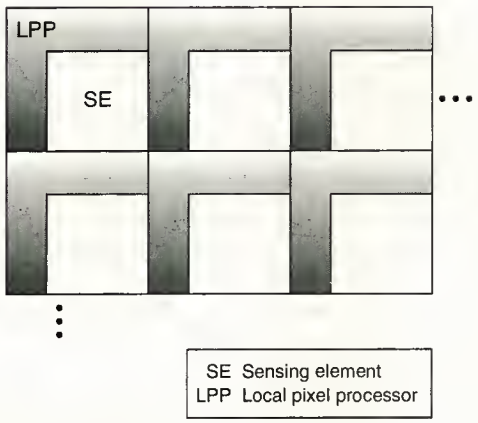


Figure 3. Sensor geometry in the focal plane.

Figure 4 shows a 2 × 3-pixel subfield. (Horizontal lines select digital rows, and vertical lines read analog data.)

Figure 5 shows the different transfer functions of a CCD (charge-coupled device) camera compared to that of an HDRC. Note that the input dynamic range that can be processed without saturation is much larger if the output signal follows a logarithmic function of the input. In Figure 5, the input signal can change its value over six orders of magnitude without saturating the HDRC device output. (That corresponds to a thermometer with a scale from 1°C to 1,000,000°C.)

The modulation of quantities like irradiated power in the space and time domains and the resolution of ratios of quantities between different pixels are even more important for image processing than the range of detected light intensities. Resolution (in the contrast and in the time and space domains) is the measure for the image quality.

The value of the above-mentioned thermometer depends on how many scale partitions one can distinguish from each

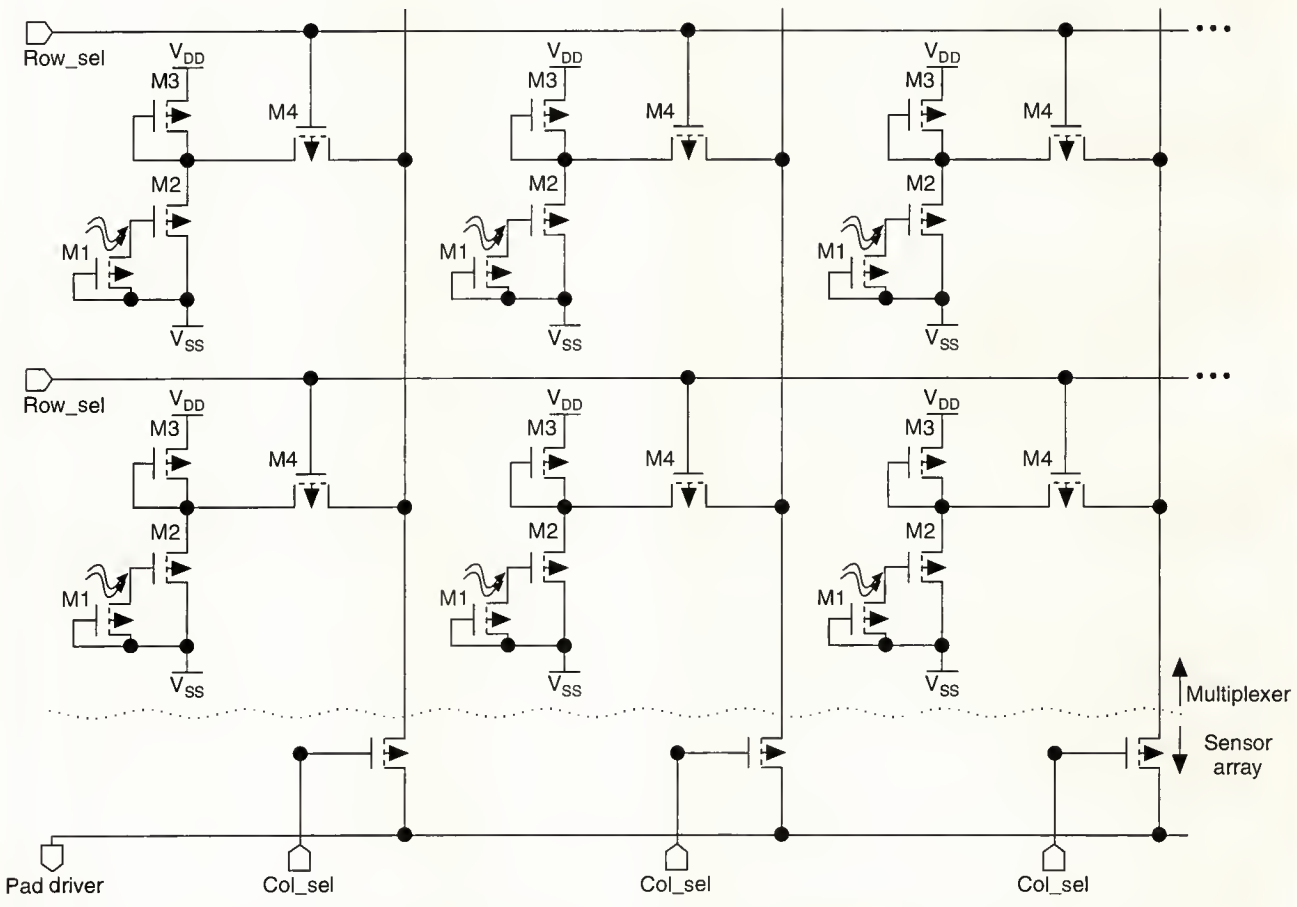


Figure 4. Circuit schematic for 3 × 2 pixels.

other; for example, whether or not you can distinguish a temperature of 100°C and 1,000°C. The thermometer's success depends on how fast it can change its value, that is, if it can react to a heat pulse within a few seconds. Thus, quality depends on the application to be met. For example, consider a high-definition video image that contains 2 megapixels, is resolved with 256 gray levels, and allows a frame rate of 100 Hz. This image, while of good quality for television applications, is not suited for high-speed imaging: The frame acquisition time is 10 ms. Also, highly dynamic scenes with contrasts exceeding a range of 1:1,000 will not be resolvable, but gray-level resolution within a given range of 1:200 (which may be displayed on TV monitors) will be superior. On the other hand, a logarithmic sensor that is optimized to handle extreme illumination conditions at the same time may not be able to resolve as many gray levels within a given range of intensities as its linear counterpart.

Figure 6 compares the contrast resolution capabilities of competing imaging systems (human eye, HDRC, and CCD camera). It is obvious that the CCD camera resolves even smaller contrasts than human eyes (at least under certain conditions). But it falls short when resolvable intensities within one scene exceed a ratio of 256:1 up to 1,024:1 (depending on the analog-to-digital converter that can be used).

HDRC imaging is thus a solution for all applications in which high contrasts must be detected at a high speed and contrast resolution of greater than 10 percent meets system requirements.

An HDRC implementation

We first integrated an HDRC chip with 64 × 64 pixels using a standard "digital" 1.2-µm CMOS technology.

Readout frequency, pixel pitch, and array size are the correlated design parameters. We chose the small array size with a medium spatial resolution (pixel pitch equals 54 µm) to get a high readout frequency. (Delay from address valid to output valid for a random access is 150 ns.)

HDRC application

The Institut de Recherches Robert Bosch SA built an experimental camera incorporating the HDRC chip, and we interfaced it to an ITEX frame grabber board for demonstration purposes. Figures 7 and 8 (next page) show the attempts to record a critical road scene using a standard CCD video camera in comparison to using the HDRC.

The scene shows two cars meeting at a tunnel's entrance. (The left car approaches the tunnel coming out of a bright zone; the right car leaves the dark tunnel region. We placed the observing camera outside the tunnel, pointing into it.) For better comparison, we extracted a zoom window of only 64 × 64 pixels corresponding to the 64 × 64 pixels of the HDRC from a standard CCD video stream. The images from the HDRC were taken with a constant aperture setting, while

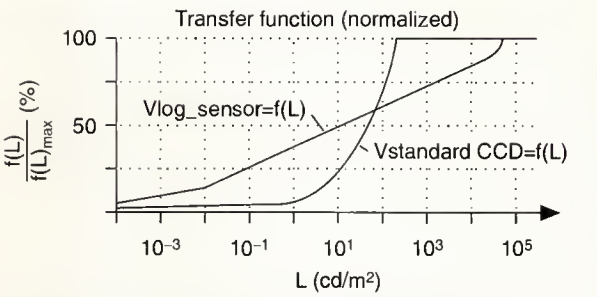


Figure 5. Transfer functions of HDRC and CCD cameras.

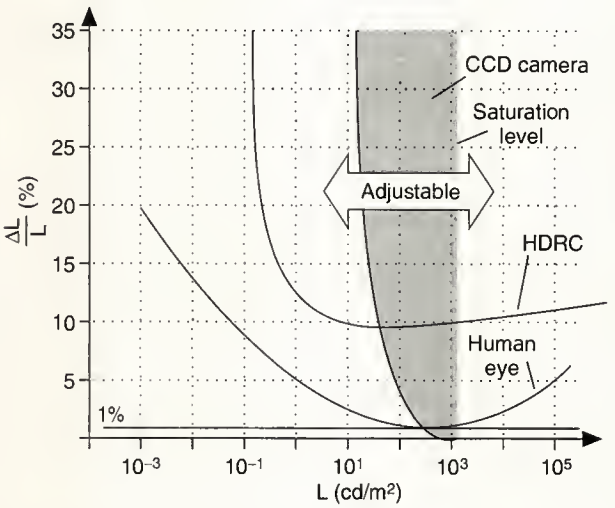


Figure 6. Contrast resolution capabilities.

the aperture of the CCD camera was set to a value that allows most details to be detected. Despite the low spatial resolution of the present HDRC, details of the cars can be extracted both in the dim and the bright regions.

In dynamic driving situations demanding short response times, the steering time for the CCD's aperture would lead to even more information loss within images taken with the CCD camera. The benefit from application of the HDRC chip in these situations is obvious and seems to be a necessary enhancement to existing vision systems in automotive applications.

Discussion

The actual 64 × 64-pixel approach with integrated digital decoders and analog output drivers is certainly not the final "production camera" for high-speed, highly dynamic imaging systems. But it proves the functionality, and it indicates the system performance of a highly dynamic range camera feasible in today's or tomorrow's standard technologies.

Integration of a complete "microsystem" with imager, decoder, and control logic in a standard CMOS process is possible today.[7] Integration of analog-to-digital converters in a digital environment is also a state-of-the-art technique.[8] Only the used die sizes limit integration of additional digital postprocessing circuitry on chip. Spatial resolution may be increased using the same 1.2-μm CMOS technology (with no space left for digital postprocessing) and the same pixel design. The design will benefit from further scaling in CMOS technology as the factors limiting the resolution are dimensions of metal width and space.

For best system performance of an image processor, an application-specific imager solution may take system requirements into account.[9] Frame rates of above 2,000 frames/second can't be reached with a single large pixel frame but are possible by partitioning the total image frame into several subfields on one chip with a parallel readout of multiple fields.

High sensitivity (below 0.1 lux) and high gray-level resolution (greater than 8 bits) may not be reached in combination with the highest spatial resolution in planar technologies; but it is possible, if one can afford a lower spatial resolution.

Still the costs for application-specific optical integrated circuits are high, because so far there is no technology-independent support for optical standard cells. This means that every optical device must be a full-custom design. Developments in recent years show that the growing market for optical solutions will need application-specific optical ICs to overcome the problems resulting from the concentration of development efforts for image sensors (within the last 20 years) on the one and only consumer application, the video camera.

Figure 7. Road scenes taken with the HDRC.

FURTHER WORK ON HDRCs WILL FOCUS on higher spatial resolution (development of an HDRC 256 × 128 chip) as well as higher contrast resolution. New functions, such as variable conversion characteristcs or active resolution control, will take even more system aspects into account. The fact that CMOS image sensors are easy to integrate will become one major aspect in the development of vision systems. All optimizations will focus on higher system performance of camera systems or image processing systems rather than toward a singular high-performance camera chip, which could be done better in other technologies than CMOS. Therefore, our work will always be embedded in the development of application-specific image processing systems.

Acknowledgments

We thank B. Ulmer, who accompanied our project as project leader at our sponsor company (Daimler Benz AG), for his contributions concerning the specification of the HDRC and his encouragement. We also thank J.F. Longchamp and R.

Figure 8. Road scenes taken with the CCD camera.

Cochard (at the Institut de Recherches Robert Bosch S.A., Lonay, Switzerland) for their very helpful conversations on image sensor development. Their work designing and manufacturing an experimental camera made it possible to demonstrate the performance of our HDRC chip. The staff at IMS helped with discussions on circuit simulation and design as well as with processing work for the HDRC chip.

The Bundesministerium für Forschung und Technologie (BMFT), Daimler Benz AG, and Volkswagen AG supported this work under contract TV8926 3. We alone are responsible for the contents.

References

1. B. Höfflinger, "Cooperative Research on Application-Specific Microelectronics," *Proc. IEEE*, IEEE Press, Piscataway, N.J., Vol. 77, No. 9, Sept. 1989, pp. 1390-1395.
2. R. Ginosar and Y.Y. Zeevi, "Adaptive Sensitivity/Intelligent Scan Image Sensor Chips," *Proc. SPIE*, Vol. 1001, 1988, pp. 462-468.
3. Y. Matsunaga et al., "A High-Sensitivity MOS Photo-Transistor for Area Image Sensor," *IEEE Trans. Electron Devices*, Vol. 38, No. 5, May 1991, pp. 1044-1047.
4. J. Heynecek, "A New Device Architecture Suitable for High-Resolution and High-Performance Image Sensors," *IEEE Trans. Electron Devices*, Vol. 35, No. 5, May 1988, pp. 646-652.
5. U. Seger et al., "Development of an Image Sensor with Enhanced Dynamic Range in Standard CMOS-Technology," *Proc. Second Prometheus Workshop*, Vol. II, Institute for Microelectronics Stuttgart, Germany, 1989, p. 371.
6. S.G. Chamberlain and J. Lee, "A Novel Wide Dynamic Range Silicon Photodetector and Linear Imaging Array," *IEEE J. State Circuits*, Vol. SC-19, No. 1, Feb. 1984, pp. 41-48.
7. D. Renshaw et al., "ASIC Image Sensors," *Proc. IEEE Custom Integrated Circuits Conf.*, 1990, pp. 7.3.1-7.3.4.
8. "IMS Semicustom ASICs: Gate Forest Family," Institute for Microelectronics Stuttgart, 1992.
9. J. Wyatt et al., "The First Two Years of the MIT Vision Chip Project," private communication.

Ulrich Seger is a staff member of the Microsystem Division of the Institute for Microelectronics Stuttgart, where he works on CMOS microsystems involving optical sensors. Earlier, he worked as a development engineer with Computer Gesellschaft Konstanz mbH. Seger studied electrical engineering at the Fachhochschule Konstanz and received the diploma in engineering for his work on a digital image preprocessor for optical character recognition.

Heinz-Gerd Graf heads the Microsystem Department of the Institute for Microelectronics Stuttgart. He has been a member of the research staff of the Department of Electrical Engineering at the University of Dortmund and a researcher in the Advanced Physics Group of Messerschmitt-Boelkow-Blohm GmbH. He received the diploma in physics from the University of Dortmund.

Marc E. Landgraf was involved in the first HDRC design as a researcher in the IMS Interface Group. Today, he is a development engineer in the Flash Memory Development Group at Intel Corporation in Folsom, California. Landgraf received his BS degree from California Polytechnic State University, San Louis Obispo, and his MS degree in electrical engineering from the University of California, Davis.

Direct questions concerning this article to Ulrich Seger, Institute for Microelectronics Stuttgart, Allmandring 30A, D 7000 Stuttgart 80, Germany; or e-mail at seger@mikroelektronik.uni-stuttgart.dbp.de.

Reader Interest Survey

Indicate your interest in this article by circling the appropriate number on the Reader Service Card.

Low 162 Medium 163 High 164

Neurocontrol for Lateral Vehicle Guidance

The complex parameterization and the nonlinear system dynamics of vehicles make the development of a controller by conventional system-theoretical methods difficult. Furthermore, this effort must be spent by experts and be repeated for each new kind of vehicle. We propose a novel solution toward autonomous lateral vehicle guidance using a neurocontroller. Neural networks can learn from measured human-driving data without knowledge of the physical car parameters. We have successfully simulated and tested this approach using an autonomous vehicle (optically steered car) on public highways.

Stefan Neusser

Jos Nijhuis

Lambert Spaanenburg

Bernd Hoefflinger

Institute for Microelectronics Stuttgart

Uwe Franke

Hans Fritz

Daimler-Benz AG

In 1986, the European automotive industry initiated the Eureka program Prometheus (Program for European Traffic with Highest Efficiency and Unprecedented Safety). It aims to collectively develop before the year 2010 an infrastructure that would reduce the

- number of accidents per driven kilometer by 30 percent,
- travel time by 20 percent, and
- traffic-related environmental damage by 50 percent, assuming an increase in traffic density by 40 percent.

In September 1991 the first results were shown to the public on the Fiat testing grounds in Turin, Italy, as CEDs (common European demonstrators).

Autonomous vehicle guidance is one of the tasks that might be realized to increase safety and efficiency of future traffic. The Prometheus subproject PRO-GEN developed this so-called copilot function using image-processing techniques supported by extensive expert knowledge engineering. The VITA vehicle is an early example.[1]

The copilot covers a wide functionality, such as collision avoidance, lane switching, and convoy driving. A basic feature is lateral control of the car, which provides a safety resource for situations in which correct driver behavior is no longer guaranteed due to tiredness or sudden illness.

Conventional controller design methods have a disadvantage in that they require an accurate model of the vehicle; furthermore, most of them are restricted to linear systems. Unfortunately, the system dynamics of vehicles show a highly nonlinear behavior with respect to velocity. One solution to overcome this problem is gain-scheduling linearization.[2]

Another solution uses neural data processing, as this paradigm implies nonlinearity in a natural way. In addition, a neural net easily adapts to the peculiar habits of each individual driver.

Literature shows a number of attempts to cover the lateral car control task by neural techniques using simulated vehicle/road systems in which car dynamics and environmental influences are grossly simplified.[3] Recently, one neural approach used a realistic car model.[4] We will take an alternative approach and capture not the vehicle characteristics but the way the car is being handled: the control task itself. Using about 50,000 of the steering actions recorded for a "flawless" human driver on a public German highway, we captured the con-

0272-1732/93/0200-0057$03.00 © 1993 IEEE

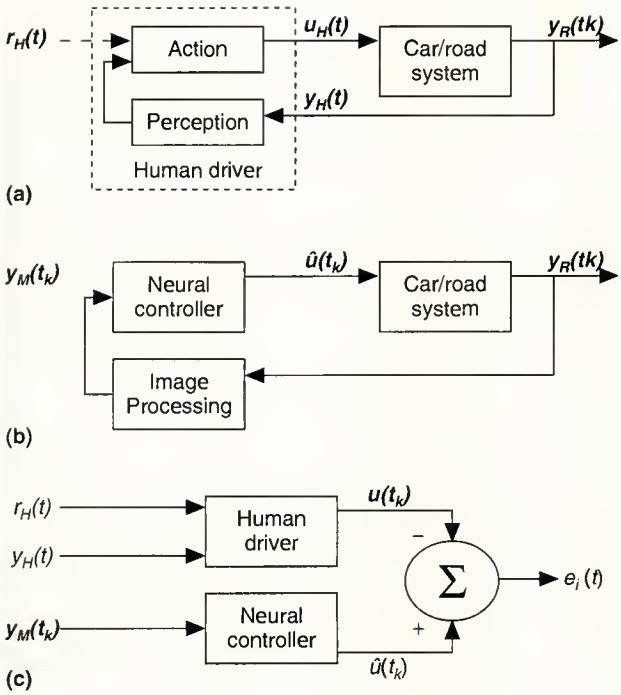


Figure 1. The general human driver/car control system with $r_H(t)$ the goal or reference inputs of the human driver, $u_H(t)$ the actuating value for the car to meet the goal inputs, and $y_H(t)$ the feedback output of the car/road system and input for the human driver (a). The neural network/car control system (b) and an identification model of the human driving behavior showing the formulation of output error (c). Only the bold variables are known.

trol task in a neural net. Subsequently, we selected a three-layer feedforward neural network with 21 neurons for this model. Finally, we installed it in the Daimler-Benz Oscar (optically steered car) vehicle for practical validation and present the first results here.

The closed-loop system

From a system-theoretical point of view, the human driver, the car, and the road form a closed-loop control system. Figure 1a generally represents this system. The actual dimension of the situation-dependent reference input $r_H(t)$ and the car/road system output $y_H(t)$ as perceived by a human are unknown. Most published efforts regarding the human steering behavior assume $r_H(t)$ to be equal to a zero lateral offset.[3-5] However, the goal directives of the human driver depend strongly on the current traffic situation (staying in a lane, overtaking another vehicle, keeping a safe distance) and will be more in line with vague linguistic statements like: "trying to keep the car on the road" and "trying to optimize driving comfort."

Figure 1b depicts the suggested neural control system. The data the neural network accepts are limited to the information delivered by the image processing system mounted on the car. The image processing system, designed by Daimler-Benz, is based on a small transputer network. The basic algorithms running on this system are described elsewhere,[6] and its reliability has been proven in combination with conventional state space controllers.[1]

In our case, the number of used outputs of the image processing system (and thus the possible number of feedback signals) is five, namely car speed $\upsilon(t_k)$, car yaw angle $\phi_{yaw}(t_k)$, road curvature $c_{ROAD}(t_k)$, road width $y_{ROAD}(t_k)$, and the lateral deviation of the car on the road $y_{OFF}(t_k)$. The car yaw angle is the angle between the car direction and the road direction. The lateral deviation (or offset) is the distance between the car's position and the road's center line. The measured output $y_M(t_k)$ is assumed to be the neural function of these five sensory signals, although sensor uncertainty and quantization noise limit the data collection quality. The real-time image processing system evaluates 12.5 images per second and computes the relevant parameters in less than 80 ms.

In our experiments, we concentrated on the basic task of staying in a lane. Therefore, the reference input to the neural controller is not explicitly necessary, but the control target is implicitly encoded in the internal net data. Obviously, the angle of the car's steering wheel is used to effectuate the lateral deviation of the car, that is, $\hat{u}(t_k) = \phi_{SW}(t_k)$; SW indicates the steering wheel angle.

Looking at both Figure 1a and Figure 1b, one can easily conclude that the problem of implementing driving behavior by a neural network can be treated as a system identification problem.[7] However, in contrast to general control approaches stated in literature, we do not identify the system to be controlled (the plant) but learn the closed-loop control task. Then, unlike normal system identification, object and model have different inputs (see Figure 1c). Equation 1 describes the assumed human driver's action: controlling only the steering angle.

$$u(t_k) = P_{HUMAN}[\, y_H(t), r_H(t)] \tag{1}$$

The objective of the neural network is to determine a function P_{NEURAL} such, that $\forall t_k \in (0,N)$:

$$\| u(t_k) - \hat{u}(t_k) \| = \| P_{HUMAN}[\, y_H(t),\, r_H(t)] - P_{NEURAL}[\, y_M t_k)] \| \leq \epsilon \tag{2}$$

for some desired $\epsilon > 0$. The values for $y_H(t)$ and $r_H(t)$ remain unknown as only $u(t_k)$ and $y_M(t_k)$ are recorded. Note that Equation 2 is a sufficient condition as long as the task is to imitate the human driver. For a stable controller, however, it

is only a necessary one. In addition, one has at least to require that the net delivers an unbiased output $u(t_k)$. All networks trained during the entire development process have met this condition.

NNSIM development environment

No detailed analytical model exists yet of the neural data processing's main advantage, identifying a plant by learning from samples. Identification requires a development environment that supports a neural network to be (re-)trained and optimized in one session without loss of data consistency. For the same reason, no loss of data consistency due to manual intervention (for instance, by file editing) can be allowed when moving the network to different realizations. This basic philosophy underlies the Neural Network Simulator.[8] NNSIM supports incremental construction, modification, and execution in automatic, interactive, and interruptible modes of operation. Of special significance is the interruptible mode, as it permits on-line changes during experimental design probing in areas where the dimensionality of the problem is not known beforehand. A global overview of the NNSIM architecture is pictured in Figure 2.

NNSIM begets its flexibility from a modular, layered software architecture, in which functionality can be enhanced incrementally by adding new functions to the procedural interface. The nature of the medium, in which the internal database is implemented (a single processor, multiprocessor, or special-purpose neural hardware), is masked by the network handler. The network handler implements the physical layer of the database and supports the construction and initialization of a network as well as simulation and import/export to other platforms. The respective procedural interfaces offer a conceptual layer to the designer. Requests, made by the standardized procedures in each procedural interface, are translated by the network handler into actions on the internal database. Therefore, every application can be created without detailed knowledge of the actual NNSIM database construction and freely moved across the various supported hardware platforms such as workstations (NNSIM_WS), personal computers (NNSIM_PC), or ASIC-based printed circuit boards (NNSIM_PCB).[9]

The menu-driven user interface offers a rich set of standard observations that have direct access to the database. This feature enhances the speed of interactive usage and does not compromise the database integrity, as observations only read the current network status. On the other hand, application-specific observations are usually guided over the standardized database access procedures. This process mainly allows a fast and secure project start without detailed knowledge of the actual NNSIM database construction. Figure 3, next page, pictures a typical NNSIM screen with a number of standard obser-

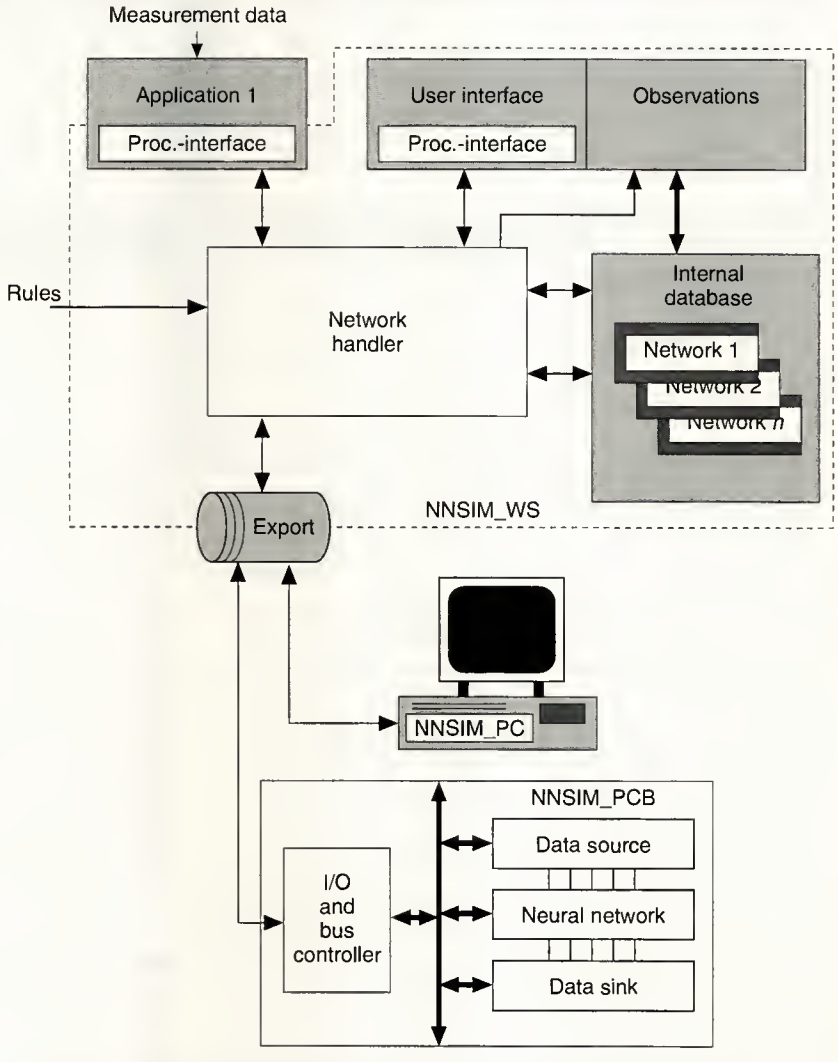


Figure 2. Architecture of the NNSIM development environment with links to PC- and PCB-level client applications.

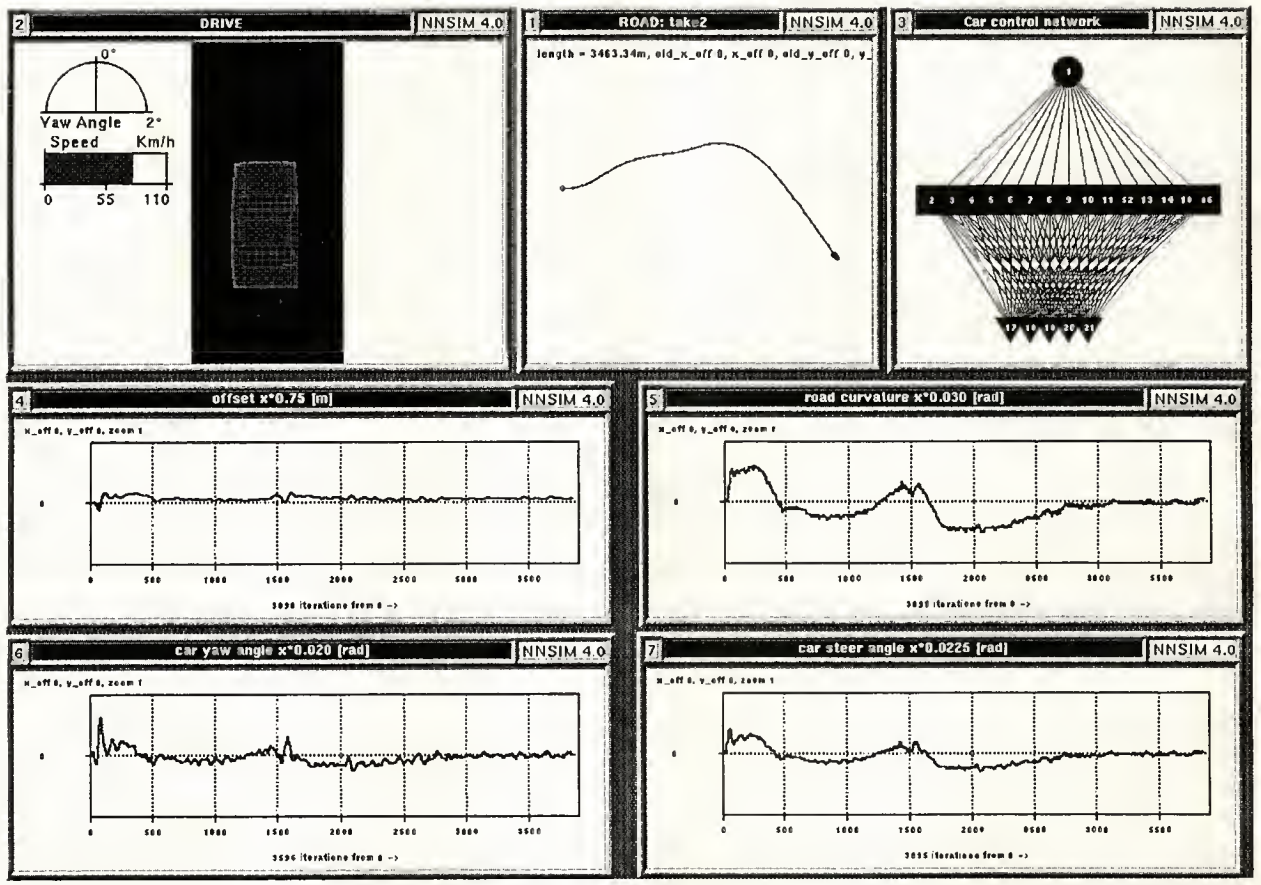


Figure 3. Typical NNSIM screen during the development of a neurocontroller.

vations and the application-specific Drive and Road windows.

In a typical design flow, the neurocontrol functionality is captured and matured at the workstation level (NNSIM_WS). If preknowledge exists, alphanumeric or graphical expert rules can initialize the network. The learning is completed from actual measurements. Artificial data are required to test the behavior of the controller in extreme situations, as these are not normally part of the measurement set.

For a first prototype test, the network is tabularized and moved to the test site, where it is included in the client application software and additional in-product fine-tuning to compensate for production spread can be performed. When the need arises, the tables can be moved back to the NNSIM environment for remedial inspection. When the network has been found to operate satisfactorily but needs further integration for reason of size and/or speed, the tables will again be returned and one or more Joplin ASICs with this same functionality are generated. This Joplin line provides digital realizations using pulse-coding techniques or Digilog arithmetic.[10]

As yet, no formal technique to prove neural functionality exists. Furthermore, neural nets are not easy to interpret; hence, there is generally a lack of confidence in the quality of a neural solution after training. We can partially solve this dilemma by providing a printed version of the neural knowledge, preferably in terms of expert rules. However, even small neural nets can comprise a vast knowledge base, which in turn leads to an extensive set of rules that must withstand thorough human inspection. Further work is required to provide a degree of structuring that enhances the transparency of the expert base.

Designing the neurocontroller

Several data sets from different human drivers have been available to train the neural network and validate its performance. They consist of 1,750 to 6,356 measurements recorded on a German federal highway with a total driving time of 140-580 seconds. In the first investigations all measured data are scaled to the range [0,1]. These initial experiments are based

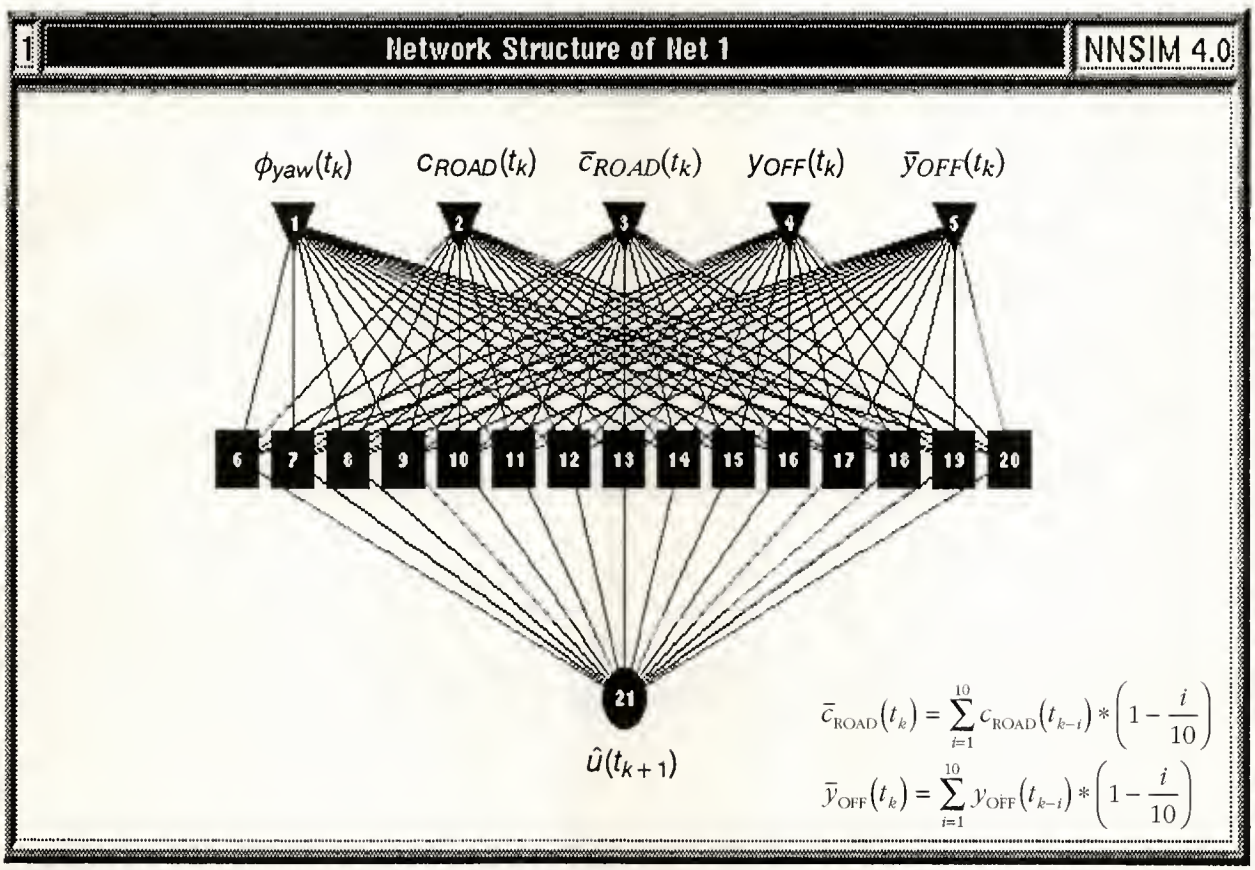


Figure 4. A small-size neural network with one output, 15 hidden neurons, and five input neurons. The network is fed with the car's yaw angle, road curvature, lateral deviation, and the time averages of both road curvature and lateral deviation.

on net structures with one hidden layer, as literature contains theoretical proof that these feedforward networks are capable of implementing any bounded continuous function $f: \mathbf{R}^n \rightarrow \mathbf{R}^m$.[11] All neurons use the sigmoid transfer function of Equation 3. The classical error back-propagation algorithm adjusts the neuron weights with the learning rate and the momentum term taken at 0.7 and 0.5 to provide a reasonable compromise between stability and speed of training.[12]

$$o_i = 1/[1 + e^{-(a_i + b_i)}] \text{ with } a_i = \Sigma\, w_{ij}\, o_j \qquad (3)$$

During these first, interruptible simulations, one can observe some correlation between the input data. A large neural net containing 50 input neurons and 135 hidden neurons learns to approximate human steering behavior. The input data contains all five measured quantities and their (up to 10) delayed values. After 100,000 learning cycles the neural net reproduces human driving actions with an average error of less than 1 percent and hardly a larger maximum error. Performing input component analysis by varying one quantity and keeping all others constant reveals some of the actual knowledge encoded in weight space. As expected, the actual output depends strongly on the road curvature, the lateral deviation, and the yaw angle of the car. Variations in car speed and lane width produce contrary effects to experienced driver knowledge, so they can be left out; their representation in the condensed learning set does not adequately reflect the physical dependencies.

With this preknowledge, the designer reduces the topology of the neural net in a second step to five input neurons and 15 neurons in one hidden layer (Figure 4). The input neurons correspond to the data signal yaw angle, road curvature, lateral deviation, and weighted time averages of road curvature and lateral deviation. This temporal memory ensures that the dynamic behavior of the vehicle can be taken into account by the net.[13] The net converges within 50,000 learning cycles to a

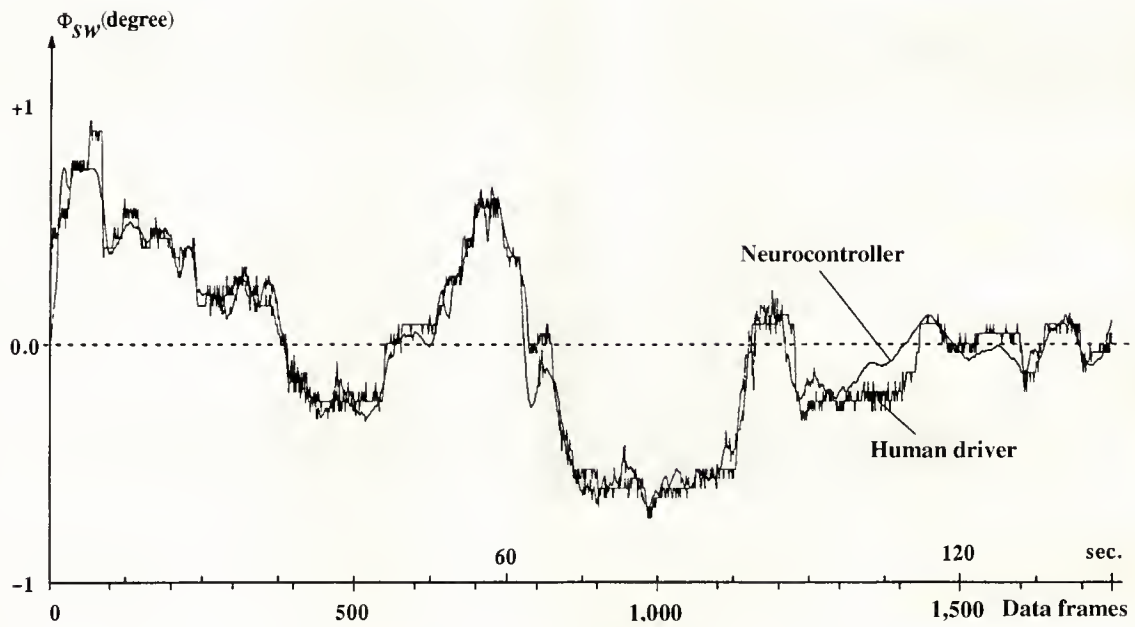


Figure 5. The steering behavior of the human driver compared with the steering behavior of the neurocontroller in Figure 4.

stable solution, whereby the input neurons are fed with one frame of measured samples during each learning cycle.

This reduced neurocontroller approximates human driver actions with an averaged error of 5 percent (see Figure 5). Looking at some parts of this diagram, we can see distinct differences in steering behavior, which might drive the vehicle off the road. Therefore the lateral control capabilities of all trained neural nets need to be investigated in a closed-loop simulation with the neural network simulator NNSIM,[8] a vehicle, and a road model. It turns out that the net uses the curvature $c_{ROAD}(t_k)$ as command variable and $y_{OFF}(t_k)$ as the variable to be controlled by the feedback loop. The values $\phi_{yaw}(t_k)$, $\overline{c}_{ROAD}(t_k)$ and $\overline{y}_{OFF}(t_k)$ control the dynamics of the car and dampen oscillations.

Experiments with different initial weight settings, numbers of hidden neurons, and human data sets indicate that the solution space of the network parameters is more sensitive to the learning set than dependent on the topology. But all potential solutions show an asymmetric behavior regarding left- and right-side offsets. Due to the scaling onto the interval between 0 and 1, 0 represents the maximum left offset. Multiplied with a static weight, 0 or values around 0 have no strong inhibiting or exciting influence on the neuron activity sum (see Equation 3). A second reason is that in the training phase 0 input values prevent weight modification. Therefore the net learns right-offset deviations or curves more extensively than left ones.

Thus, in a third step we chose a symmetric output function as described by Equation 4. The standard sigmoid function is scaled and shifted to yield outputs in the range [−1,1]. This kind of function overcomes the problem just discussed.

$$o_i = 2/[1 + e^{-(a_i + b_i)}] - 1 \text{ with } a_i = \Sigma\, w_{ij}\, o_j \qquad (4)$$

To make use of the full value range, we additionally replaced the sigmoid output neuron with a semilinear neuron with saturation points at −1 and 1. Figure 6a–c shows the driving behavior of the new small-size neurocontroller on a simulated road. The neurocontroller keeps the car on the road within 0.16 meters, peak-to-peak offset drift. Like the human driver, the net shows a static offset of 0.17 meters to the right-hand side. Further simulations on extreme situations reveal that the generalization capability of the net lets the controller handle offset deviations and curvatures much larger than those included in the learning set.

Simulations and experimental results

As stated earlier, the problem of lateral vehicle guidance has also been investigated using conventional PID controllers.[4] We therefore compared the performance of the neurocontroller with that of classical ones. For this reason, the trained controller as well as a conventional approach are simulated in a closed loop using a simplified model of the

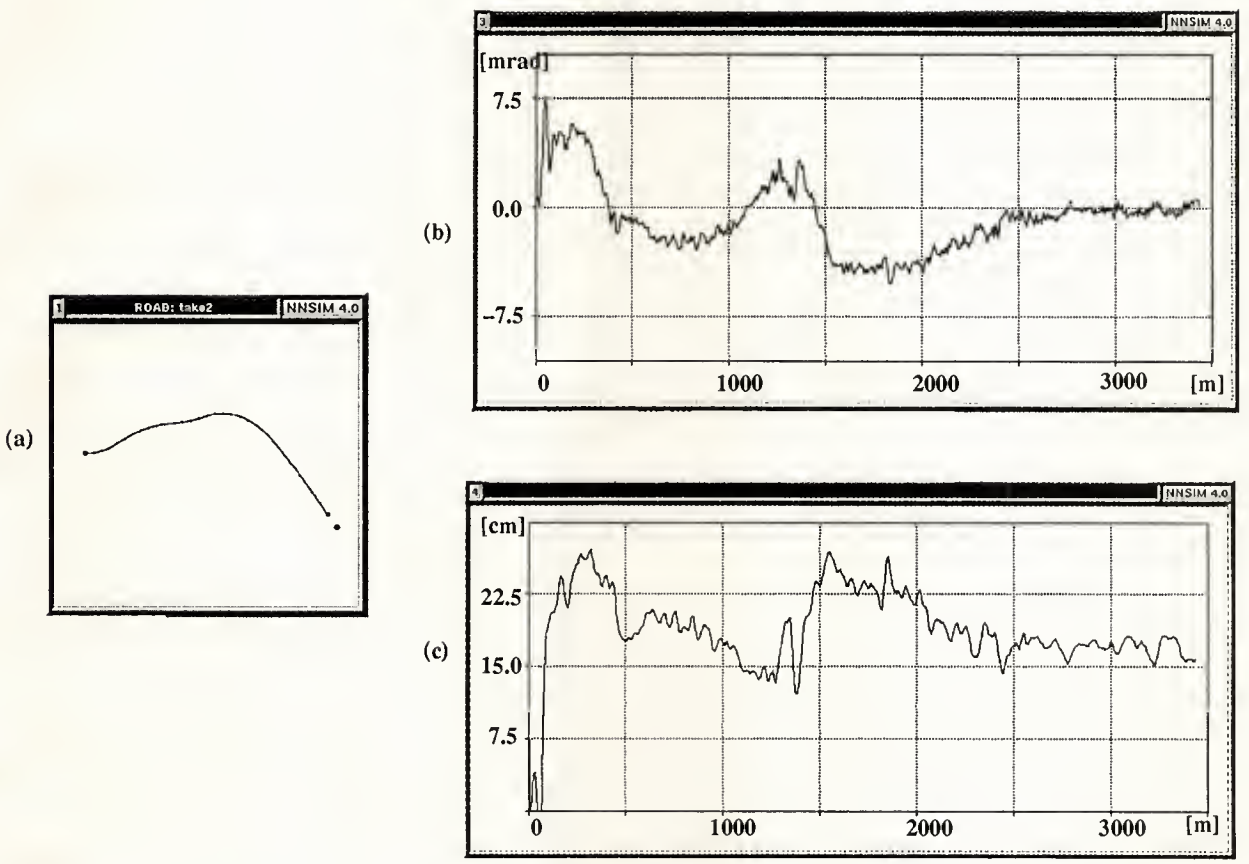


Figure 6. Simulated driving behavior of the small-size neurocontroller at 80 km/h: part of a simulated highway (3,460-meter length) seen from a bird's eye view (a), car steering angle (b), and lateral deviation of the car (c).

lateral vehicle dynamics. Figure 7a, next page, depicts the step response of the neural system for a lateral offset of 1 meter. The conventional linear state controller has the following four state variables: offset $y_{OFF}(t_k)$, yaw angle $\phi_{yaw}(t_k)$, yaw angle velocity $d[\phi_{yaw}(t_k)]/dt$, and steering angle $\phi_{SW}(t_k)$. Figure 7b gives the result of this simulation.

The speed in both simulations is 25 meters/s. The neural controller shows an aperiodic behavior with respect to the offset. The remaining constant offset of 0.17 meters is caused by the tendency of the human driver in our training set to keep to the right of the lane's center. From the large slope of the steering angle at the beginning, one can see that the controller produces a large steering angle velocity. The sharp asymmetric bend of the steering angle indicates its nonlinear nature. In contrast, the linear controller shows higher order dynamic behavior, resulting in an overshoot of the offset and a larger yaw angle. If the speed is further increased, the neural controller keeps its aperiodic behavior, whereas the linear controller tends to a larger overshoot. Since there is currently no theoretical proof for stability of a neural controller, we carried out various additional simulations with different initial conditions. In all these tests the neural controller shows a satisfactory behavior.

After this preparatory work, we performed realistic experiments with a Mercedes-Benz car (300 TE) on a public highway near Stuttgart. Figure 8 represents the results for the neural controller and the conventional controller. Both diagrams show the curvature profile $c_{ROAD}(t_k)$ of the part of the road we selected for the test, as well as the offset $y_{OFF}(t_k)$ and yaw angle $\phi_{yaw}(t_k)$. We multiplied the curvature by a factor of 100 for better visualization. Note that a curvature of 0.002 m^{-1} corresponds to a radius of 500 meters, which is not typical for a modern highway but can obviously still be encountered on the older ones. Although we attempted to keep the speed constant at 80 km/h during these tests, both diagrams differ by about 3 seconds, as can be seen from the shifted curvature profile.

The surprising fact that the offset produced by both controllers has a negative mean value is caused by a strong lateral banking of the road to the left. Since this was unknown to the controllers, it acts as a permanent disturbance. Apart from this deviation, practical experiments confirm the expectations from these simulations. The excellent results of the simulations given in Figure 6, however, cannot be reached since (on actual roads) further disturbances like cross wind, grooves in the lane, badly painted markings, and so on tend to activate the system.

Obviously the neural controller produces smaller offset variations compared to the state controller. This corresponds with smaller yaw angles of the vehicle. Calculation of the yaw angle variances yields σ_{yaw}= 0.20 degree/s for the neural controller and σ_{yaw}= 0.27 degree/s for the conventional one. This behavior results from stronger steering activity, and the difference is clearly noticeable for the passenger. In all, the neurocontroller was felt to be the most comfortable of the two.

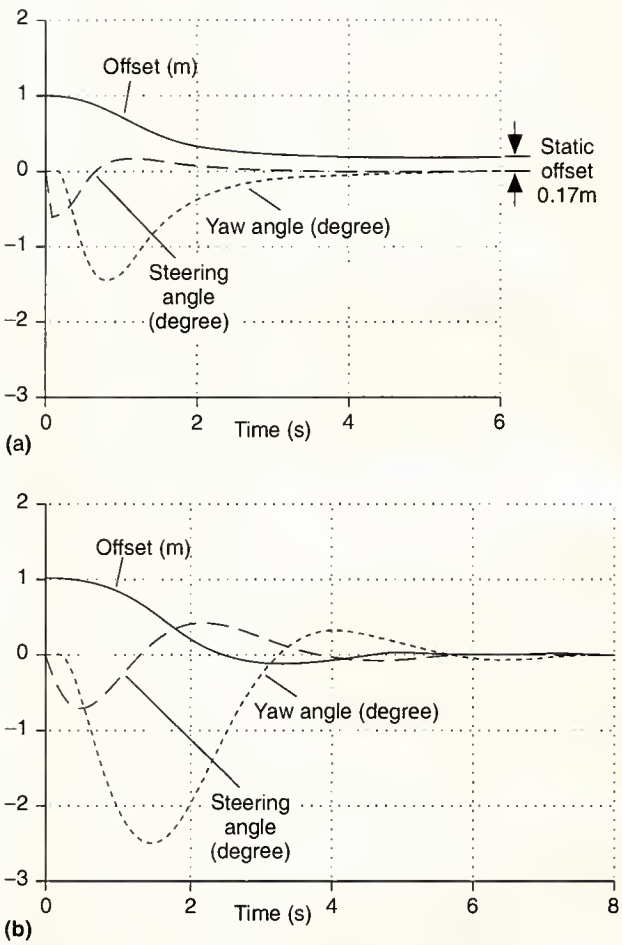


Figure 7. Step response of both investigated controllers: neural system (a) and conventional (b).

THE SIMULATIONS AND PRACTICAL TESTS WE DESCRIBED confirm that a small-size feedforward autonomous neural network (21 neurons) can learn to steer a vehicle at high speeds only from looking at human-driving examples. In this way, the network learns the total closed-loop behavior including the nonlinear dynamics of the vehicle as well as the driver's individual driving style. It stands to reason that the behavior to be learned should previously be proven to be correct as the neurocontroller will obviously not be capable of improving on its human example.

Besides the performance of a neural system versus a conventional one, the design effort for both approaches is a key question. Where the training algorithms for neural nets still consume much computation time, only a little knowledge of the underlying physical process is necessary. On the other hand, the design of a state controller requires a deep insight into the dynamics of the system. The conventional controller considered here was designed by experts in vehicle dynamics and control with years of experience, and the neurocontroller by the ultimate laymen.

An advantage of the classical design methods, which cannot be overlooked, is the existence of stability proofs that are valid as long as reality is adequately described by the used model. However, we are convinced that for small neural systems like the one considered here, stability can sufficiently be shown by exhaustive closed-loop simulations, which preaches in favor of neurocontrol.

The main result of our practical investigations is that the neural controller trained on human-driving examples exhibits an aperiodic behavior that does not vanish at higher speeds (tests performed up to 130 km/h). It produces less lateral deviations than the linear state controller and gives a pleasant driving feeling.

Acknowledgments

Daimler-Benz A.G. and the Bundesministerium fuer Forschung und Technologie supported this work under contract TV 8926 3. We gratefully acknowledge our cooperation with the partners in the Prometheus/PRO-CHIP project 23.232. Last but not least, we are indebted to B. Haneberg for his vigilance and to the many people supporting the IMS design, fabrication, and test facilities.

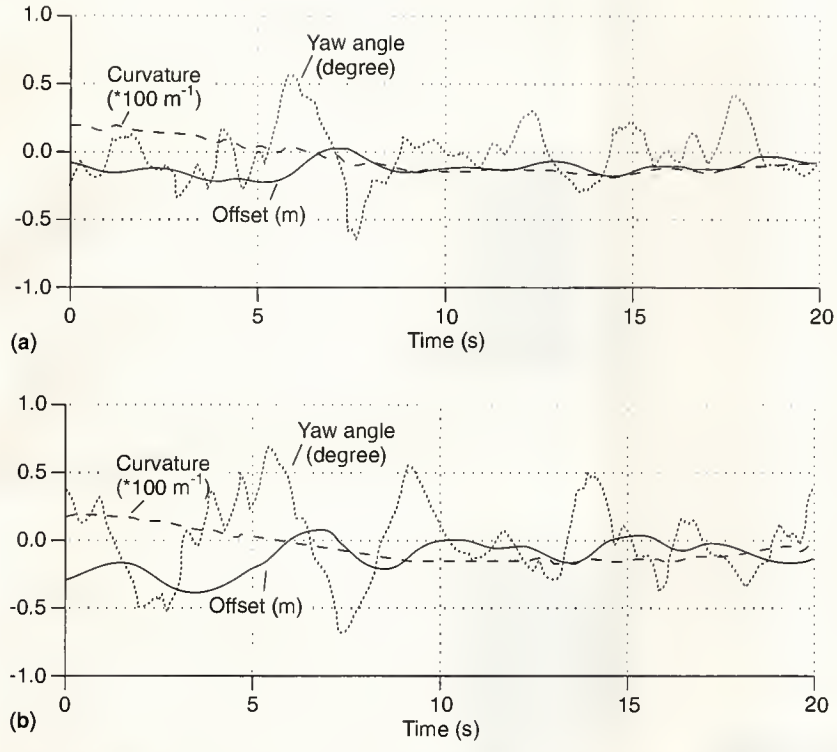


Figure 8. Experimental results for both investigated controllers: neural (a) and conventional (b).

References

1. U. Franke, H. Fritz, and S. Mehring, "Long Distance Driving with the Daimler-Benz Autonomous Vehicle VITA," *Proc. Prometheus Workshop*, Prometheus Office, Stuttgart, Germany, Dec. 12-13, 1991, pp. 239-247.
2. E.D. Dickmanns and A. Zapp, "Autonomous High Speed Road Vehicle Guidance by Computer Vision," *Proc. 10th Int'l Federation of Automotive Control (IFAC) World Congress*, June 1987, pp. 161-164.
3. J.F. Shepanski and S.A. Macy, "Manual Training Techniques of Autonomous Systems Based on Artificial Neural Networks," *Proc. IEEE First Int'l Conf. Neural Networks*, M. Caudill and C. Butler, eds., June 21-24, 1987, pp. 697-704.
4. K. Mecklenburg et al., "Neural Control of Autonomous Vehicles," *Proc. IEEE Vehicular Technology Conf.*, May 1992, pp. 205-212.
5. R.A. Hess and A. Modjtahedzadeh, "A Control Theoretic Model of Driver Steering Behavior," *IEEE Control Systems*, Vol. 10, No. 5, Aug. 1990, pp. 3-8.
6. U. Franke, "Real-Time 3D Road Modeling for Autonomous Vehicle Guidance," *Selected Papers Seventh Scandinavian Conf. Image Analysis*, P. Johanson and S. Olsen, eds., World Scientific Publishing Company, Aalborg, Germany, 1992, pp. 316-323. Publishing Company, Aalborg, Germany, 1992, pp. 316-323.
7. K.S. Narendra and K. Parthasarathy, "Identification and Control of Dynamical Systems Using Neural Networks," *IEEE Trans. Neural Networks*, Vol. 1, No. 1, Mar. 1990, pp. 4-27.
8. J.A.G. Nijhuis et al., "NNSIM: An Environment for the Simulation of Application with Fine-Grained Parallelism," *Int'l J. Computer Simulation*, Ablex Publishing, Norwood, N.J., Vol. 2, No. 2, 1992, pp. 195-212.
9. S. Neusser et al., "Rapid Prototyping of Neural Network Applications," *IEEE Digest SPIE 92*, Mar. 1992.
10. L. Spaanenburg et al., "A Multiplier-less Digital Neural Network," *Digest Second Int'l Conf. Microelectronics of Neural Networks*, Oct. 1991, pp. 281-289.
11. K. Hornik, M. Stinchombe, and H. White, "Multilayer Feedforward Networks Are Universal Approximators," *Neural Networks*, Vol. 2, No. 5, Sept. 1989, pp. 359-366.
12. D.E. Rummelhart, G.E. Hinton, and R.J. Williams, "Learning Internal Representations by Error Propagation," *Parallel Distributed Processing*, D.E. Rummelhart and J.L. McClelland, eds., MIT Press, Cambridge, Mass., Vol. 1, 1986, pp. 318-362.
13. T. Troudet et al., "Towards Practical Control Design Using Neural Computation," *Proc. Int'l Joint Conf. Neural Networks*, Vol. II, July 8-12, 1991, pp. 675-681.

Stefan Neusser is a member of the technical staff at the Institute for Microelectronics Stuttgart in Germany. His research interests include the development and realization of fuzzy- and neurocontrollers. Neusser received the Dipl Ing degree in electrical engineering from Stuttgart University.

Jos Nijhuis currently works with the Centre for Microelectronics Eindhoven in The Netherlands and was a member of the technical staff at IMS when this article was written. Earlier, he worked as a research assistant at Twente University of Technology in The Netherlands. His area of research includes the design and analysis of artificial neural networks including ASIC realizations. Nijhuis received the Dipl Ing degree in electrical engineering from Twente University of Technology.

Lambert Spaanenburg is a senior scientist at IMS and earlier served as an associate professor in the Department of Electrical Engineering of Twente University of Technology. His interests include design and test principles of digital and neural systems. Spaanenburg has authored over 80 technical papers, served on several editorial boards, and regularly consulted to industry.

Bernd Hoefflinger's biography and picture appear on p. 10 in the Guest Editor's Introduction.

Uwe Franke is a scientist at the Mercedes-Benz Research Institute in Stuttgart and previously worked as a research assistant at the Institute for Communication Engineering in Aachen, Germany. His main interests include computer vision, particularly for autonomous vehicle guidance, aspects of parallel computing, and artificial neural nets. Franke received the Dipl Ing degree in electrical engineering and the PhD degree from the Technical University Aachen.

Hans Fritz is a member of the technical staff at the Mercedes-Benz Research Institute in Stuttgart. Previously, he worked as a research assistant at the Institute of Systemdynamics and Control in Stuttgart. His main area of interest covers longitudinal and lateral control of autonomous road vehicles using computer vision techniques. Fritz received the Dipl Ing degree in electrical engineering from the Technical University Karlsruhe, Germany.

Direct any questions concerning this article to Stefan Neusser, Institute for Microelectronics Stuttgart, Allmandring 30A, 7000 Stuttgart-80, Germany; or e-mail at neusser@mikroelektronik.uni-stuttgart.dbp.de.

Reader Interest Survey

Indicate your interest in this article by circling the appropriate number on the Reader Service Card.

Low 165 Medium 166 High 167

Special Report:
Supercomputing–the View from Japan

Information technology is the most important area of research in Japan, with industry spending the bulk of the funds, followed by private and government research institutes and universities. To keep pace with developing information processing needs, MITI's Superspeed project investigated high-speed novel devices and computer architecture, algorithms, and languages for parallel computing.

David Kahaner

US Office of Naval Research

[David Kahaner is on assignment with the US Office of Naval Research. He generally comments on activities in the Far East for inclusion in the Software Report column. Since we felt readers would be interested in a detailed description of supercomputing trends in Japan, we also offer this special report. His comments are his own; they do not express any official policy.–Ed.]

In Japan, information technology is the most important area of research besides life sciences and environmental research. For example, Japan's 1989 research and development budget for information processing was about ¥1,012 billion, of which ¥958 billion were spent by industry, ¥24 billion by private research institutes, ¥23 billion by universities, and ¥5 billion by governmental research institutes. (There are approximately 125 yen per US dollar.)

Superspeed project

At the end of the 1970s, as it became apparent that future information processing needs would require new computer architectures and new devices, Japan's Ministry of International Trade and Industry (MITI) went the usual way in bringing together experts from universities, governmental research laboratories, and industry to formulate a project proposal. The outcome was quite unusual, however, as MITI decided to run two large projects in parallel; the High Speed Computing System for Scientific and Technological Uses Project, dubbed the Superspeed Project, (1981 to 1989, ¥23 billion) and the Fifth Generation Computer System Project (1982 to 1991, ¥55 billion). While the FGCS Project aimed at a risky, new computing paradigm, cutting relationships to existing computer systems, the Superspeed Project was more of an extension of the present systems. It aimed at the development of a high-speed computing system for scientific and technical applications. The target system was supposed to operate at a rate of more the 10 Gflops, which was 100 to 1,000 times faster than the speed of conventional computers at that time. Two major R&D projects were conducted: one on high-speed novel devices and one on computer architecture, algorithms, and languages for parallel computing.

The six major vertically integrated computer/semiconductor companies–Fujitsu, Hitachi, Mitsubishi, NEC, Oki, Toshiba–together with the Electrotechnical Laboratory (ETL) participated in the project. The research on high-speed devices was divided among the six participating firms: NEC, Toshiba, Hitachi, and Mitsubishi researched GaAs chips; Fujitsu, Hitachi, and NEC, Josephson junctions; Fujitsu and Oki, HEMT (high electron mobility transistor) devices.

The research on parallel processing was divided into three subgroups: a high-speed parallel (four-CPU) subproject (called Parallel, Hierarchical

0272-1732/93/0200-0067$03.00 © 1993 IEEE

US supercomputers have more CPUs, each having a small number of pipes. Japanese machines have had fewer CPUs, but each has more pipes.

Intelligent computer project, or PHI); the Sigma-I dataflow subproject; and a satellite image processing subproject. Of these, PHI was the most important. In a practical approach to developing a four-CPU machine as quickly as possible, the subproject combined four of Fujitsu's existing one-processor VP2000 supercomputers. To this combination was added a large high-speed common memory. Since each of the VPs already had its own memory, the concept of a hierarchical memory structure appeared. The idea was that a user should not have to know about this hierarchy and could treat the memory as "flat."

The project was concluded in 1990 by demonstrating the PHI system to the evaluation team. The prototype high-speed parallel system using four processors ran at over 10 Gflops, peak, and had real performance of over 1 Gflops. NEC wrote and tested one benchmark that solved a very large (32K) system of linear equations in under 11 hours. This was not a prototype of a machine that could be directly commercialized. GaAs devices–HEMTs and MESFETs–were used, though not as extensively as envisioned. Josephson junction devices were not used at all, though advances in such devices put Japan in the lead in this area. Less tangibly, the project focused the private sector on supercomputers at a critical time, earlier and more heavily than they would have done individually. Of course, cooperation also meant that work was done faster and more economically. Individually, the Japanese companies were also investing heavily, and some estimates were as high as three to four times the government figure, ¥300 to ¥500 million by each of the three.

Supercomputing

There are between 400 and 500 supercomputers installed worldwide (excluding IBM installations which are difficult to count); about 125 of these are now in Japan. Three large Japanese electronic companies, NEC, Fujitsu, and Hitachi, produce shared-memory supercomputers with some parallel features; these are products, and are supported and marketed as such. Within Japan, Fujitsu has almost half of the supercomputer installations, with Cray, Hitachi, and NEC sharing the balance.

There are about 40 supercomputers at Japanese universities, but the number could be misleading because at least a third are older machines or others with very modest performance. Most Japanese university scientists can get supercomputer time, but rarely on top-end machines which are mostly found at industrial labs or in the prestigious national universities. Access to supercomputers at Japanese universities has improved markedly in the past two or three years, though in my opinion, it is still below what is available to US academics.

Networking has improved recently, but academic networking is not as ubiquitous as it is in the US. The prestigious universities have excellent services, while many other universities have none. There are more high-performance networks in the US than in Japan. Network interconnectivity in the US is also much better than in Japan; several more or less independent Japanese networks are supported by different Ministries. Researchers in Japan sometimes communicate with each other or with colleagues in Europe by transiting through the US. Counterparts to very high performance networking projects in progress or planned in the US have not yet jelled in Japan. However, Japan has excellent, sometimes unique technology, including a large infrastructure in the ISDN, and their networking difficulties seem to be more social, organizational, or cultural than technological. Nevertheless, research in supercomputing trails that of the West, except for applications developers working on commercial software packages.

Architecture and performance

Today's supercomputers have large memories, 1 to 32 Gbytes, and several (currently up to 16) independent and very high performance CPUs, which are sometimes called functional units or FUs. Within each CPU are several pipelines (pipes) consisting of the components that add, multiply, and so forth. (Within a CPU the pipes have only one instruction path and must all carry out the same calculation, whereas different instructions can be executing on the independent CPUs.) A floating-point operation is not achieved until the pipe has been filled, but once this happens a new floating-point operation occurs each clock cycle (hence the term pipe). Data can be moved to and from memory at rates of up to a few gigabytes per second, but this is not fast enough to keep up with the arithmetic performance. Thus some kind of memory hierarchy is employed. For example, within each CPU, data from memory first goes to registers, which are built of the fastest and most expensive static-RAM chips and have a capacity up to about 1 Mbyte. Under certain circumstances, the pipelined arithmetic units can operate on data from the registers at the peak hardware speed.

An essential difference between US and Japanese supercomputers has been that US supercomputers have more CPUs, with each having a small number of pipes. Japanese machines have had fewer CPUs, but each has more pipes—up to 16. This situation arises mostly because US companies

have more experience building multi-CPU machines, but the distinction is slowly changing as the Japanese add more CPUs to their systems.

Peak performance can be computed from the hardware specifications of the machine. It is obtained by dividing the total number of independent add and multiply pipes by the clock cycle time in nanoseconds to produce a result in gigaflops. Performance of Japanese supercomputers is always specified in terms of the peak that the hardware can achieve. Peak performance varies from about 5 Gflops for Fujitsu's VP2600 (billions of 64-bit flops) to 32 Gflops for the Hitachi S-3800. The Cray Y-MP C90 has a peak speed of about 15 Gflops. NEC's SX-3 has a peak of 26 Gflops.

Of course, most real applications will exhibit performance far below the peak. Actual performance is measured in terms of throughput, performance on specific applications or benchmarks, and other criteria. (Informally, many scientists assume that usable speed is one order of magnitude less than the claimed peak.) This rate can be heavily influenced by how rapidly and in what quantity data can be moved around. The start-up time to fill a pipe from a register is an overhead, and it will reduce the computing speed unless it can be amortized over a sufficiently large number of calculations. If there are many pipes, subdividing arrays to use them all reduces the number using each and increases the relative importance of the start-up. Also, bandwidth between memory and registers must match the realizable speed of the CPUs. There is additional overhead (memory latency) arising in the process of fetching numbers from memory for deposit in the registers; this depends on the type of memory chips used, how skillfully irregular retrievals are carried out, and whether bank or other conflicts in memory are avoided. In real problems, there are significant fractions of the program that require floating-point computation of scalars as distinguished from arrays. Some supercomputers such as Fujitsu's VP2000 have two separate scalar arithmetic units for each CPU operating concurrently with the vector (array) unit. Like data movement, these scalar units are not relevant in computing peak performance, but are important in measuring real performance.

The key to building a high-performance supercomputer is to balance memory capability, arithmetic processor performance, data movement capability, and other components. Each component plays a crucial role. This is generally related to the overall architectural design of the system, and is an area in which Cray has been particularly strong.

Supercomputer technology

Another way to make machines faster is to use faster components, hardware, and devices, and the Japanese have excelled here. NEC states explicitly in its 1990 annual report, "... the actual performance of a supercomputer is determined by its scalar performance .... NEC's approach to supercomputer architecture is clear. Our first priority is to provide high-speed single-processor systems which have vector processing functions and are driven by the fastest technologies, while giving due consideration to ease of programming and ease of use; we also seek to provide shared memory multiprocessor systems to further improve performance." The Japanese see four major hardware tasks as being key to additional performance: faster chips, smaller size, heat reduction, and elimination of logic bugs.

Hitachi's 1992 supercomputer uses 25,000 gate arrays, NEC's (1989) has 20,000, Fujitsu's (also 1989) has 15,000.

Supercomputers from NEC, Fujitsu, and Hitachi use tried and true emitter-coupled logic (ECL) semiconductor technology for basic processor chips, but have pushed their capabilities in this area quite far. For example, clock cycle times vary from 3.2 ns (Fujitsu), to 2.5 ns (NEC), to about 2.0 ns (Hitachi). These figures are better than US products (the Cray Y-MP C90 has a cycle time of 4.2 ns). Faster clocks translate into better performance. Another example of technology advance is in the area of lithography, the process of outlining circuits. Beginning as an optical process generating 10-μm line widths in the 1960s, the practice is now an X-ray process in the 0.8- to 0.5-μm range. As line widths become narrower, more highly packed chips can be built. The Japanese are aggressively working to reduce line width, and also to improve width variability in the hopes that the former will translate into direct performance improvements and the latter into less conservative designs. ECL gate densities are also improving. Hitachi's newly announced (1992) supercomputer uses 25,000 gate arrays, NEC's (introduced in late 1989) has 20,000 gate arrays, and Fujitsu's (also introduced in 1989) uses 15,000 gate arrays.

High-end Japanese machines all have water-cooled CPUs, but slightly slower air-cooled versions are also available. In addition, air cooling is used in peripheral devices. Fujitsu uses GaAs chips in some of its peripherals so these can be effectively cooled by air (GaAs can run cooler than silicon). Generally, the use of exotic device technology has been fairly conservative, although there are research projects at all the large Japanese companies. Thus far GaAs is not being used for CPU chips in any commercial Japanese machines, nor are even more sophisticated Josephson junction circuits. Fujitsu used the Superspeed Project results to develop a hybrid Josephson junction-VLSI device, and plans to use it in its next-generation supercomputers, probably out in the mid-1990s.

There is no work in Japan on standardization of scientific software, and almost no research comparable to that in the West on portable numerical algorithms.

(It takes three to five years to produce a large-scale supercomputer product.) Similarly, NEC developed GaAs logic devices as well as memory chips and has designed a multichip package for supercomputers. GaAs is seen as slowly replacing ECL, though the Japanese are convinced that performance gains can still be obtained with silicon.

Supercomputer software

All three Japanese supercomputers now are available with a customized version of the Unix operating system. The use of Unix will help the migration of application programs onto Japanese systems. People are just now coming to grips with the need to assess software costs, and moving to Unix is clearly seen as one way to reduce costs. In Japan, this is a change from the use of proprietary operating systems that has occurred only in the past two or three years. For Hitachi it is only just now occurring, and the company has not totally embraced Unix. Its newest supercomputer is available in a Unix version, and also with the company's own IBM-like operating system for compatibility with older Hitachi systems. The situation is similar for Fujitsu, which also supports both Unix and its own system.

In the past, applications developed in the West have been installed very slowly, which was a major impediment to the purchase of Japanese supercomputers both in and outside Japan. Using Unix will improve this situation. However, using a standard operating system only means that software portability is improved and development time is reduced, not that a program will run efficiently. There does not yet seem to be any shortcut to maximum performance short of incorporating knowledge of the hardware into the algorithms and software.

Early Japanese supercomputer software development was limited to producing Japanese language interfaces for Western software products, and this is still an important activity. For example, NEC has recently moved the latest version of the heavily used engineering analysis system Nastran to its supercomputers, and the company's supercomputer promotional literature lists about 100 products (many from the West) that are available in a wide range of disciplines. Other vendors are engaged in similar projects. But more recently, first-rate packages designed and implemented in Japan are appearing. Good examples are:

- DEQSOL from Hitachi for the solution of the partial differential equations arising in engineering simulation,
- Alpha-flow from Fuji Research Institute for solution of fluid dynamics problems,
- Fortran/a from Fujitsu, allowing object-oriented programming from within a Fortran environment, and
- AMOSS from NEC for molecular orbital calculations.

For those users who need to create software (rather than using existing applications), standard languages such as Fortran and C are available on all Japanese supercomputers, and the vendors are careful to ensure that these meet all announced standards, although they have various enhancements too. To get efficient programs, users can rearrange their algorithms, insert special directives within their programs, and also use vendor-provided automatic vectorizers and autotasking. Optimized vendor libraries with simple interfaces are another good way to obtain efficiency. The three Japanese supercomputer companies have large teams of programmers developing these libraries, and they also support well-known commercial libraries from the West; IMSL and NAG, and noncommercial projects such as Eispack and Linpack, among others. If the user interfaces are standardized, portability is maintained along with efficiency. But there is no work originating in Japan with an eye toward standardization of scientific software. Also, there is almost no research comparable to that in the West on portable numerical algorithms, as typified by the Lapack project at the University of Tennessee and other cooperating places. Nor is there much pressure to develop standardized software; vendors and users still develop libraries and user interfaces for their own platforms and applications. Japanese computer users can and do write their own application software. People who have studied it from the inside claim it can be quite good.

Questions regarding this column can be addressed via e-mail to David K. Kahaner, US Office of Naval Research, Far East, at kahaner@cs.titech.ac.jp.

"Fair is foul, and foul is fair"

Stephen L. Diamond

SunSoft, Inc.

Phone (415) 336-4190

Fax (415) 336-4477

steve.diamond@eng.sun.com

Everyone seems to believe in open systems, but curiously no one seems to agree on what they are. Claims of "openness" are everywhere, and are nowhere more prevalent than in the advertisements of many computer hardware and software companies. One software vendor proclaims that its operating system product is "open," presumably because anyone can openly buy it at a local software store. A hardware vendor claims openness because its proprietary computer is based on a microprocessor chip that anyone can buy. Another hardware vendor claims that a proprietary computer architecture is "open" because it runs a proprietary operating system available from several other hardware companies, which make the same assertion. Nearly all users and many vendors are tired of these self-serving claims.

More importantly, how does the IEEE standards development program relate to such claims of openness? Let me use an older analogy to relate the standards activities of the IEEE to the current outbreak of "openness." A classic recipe with which many people are familiar is the famous scene from Macbeth (Act IV, Scene i), in which the witches chant:

> Double, double toil and trouble;
> Fire, burn; and, cauldron, bubble.
> Fillet of fenny snake,
> In the cauldron, boil and bake;
> Eye of newt, and toe of frog,
> Wool of bat, and tongue of dog.

Frequently when whipping up the souffle of openness, the same sort of recipe will be used, with a dollop of standards thrown in. This witches' brew of product attributes then produces something like "industry standards," which can—and usually does—mean just about anything. As with many vague recipes, the resulting product is usually unreproducible, since the original rules for creation didn't specify how hot the cauldron was supposed to be (bubbling temperature?), how big the fenny snake fillet (the 6- or 12-ounce variety), nor what type of dog to use.

Contrast this with the recipe for a formal standard. The rules are simple since ANSI requires that all American National Standards be developed using rules that mandate openness, fairness, and equity. Proper rules—ones that all participants help to determine—require precise definitions and specify exact amounts. Requirements are agreed upon in a public forum, and constant review makes sure that the recipe is publicly available and capable of being duplicated. The result is an accepted agreement on a way to "do something"—whether it is to create a local area network (IEEE Std. 802) or a RISC (reduced instruction-set computer) architecture (IEEE P1754). These recipes are published, not in Shakespeare nor the fiction section of a library, but in books with the title of *American National Standards* or *International Standards.* And these books are available for sale and for use in implementing a product based upon the interface specified in the standard. The standards process has been around for a long time. So why is there a sudden need to embrace "openness" and the invention of all of the types of new open recipes? Simply put, open systems are hot today because customers want them.

Open systems offer users better value and safety, allowing customers to protect their investment in the face of the increasing globalization and specialization of the information technology industry. Users of open systems are less subject to unpleasant surprises in price, performance, or

0272-1732/93/0200-0071$03.00 © 1993 IEEE

availability, at the whim of a vendor. Because today's users are decentralizing and distributing their applications across heterogeneous networks, the number of distributed applications is increasing dramatically. Such a world is greatly facilitated by truly open systems.

Another option, of course, is to make a single vendor the answer to all of your computing needs, which works equally well for ensuring interoperability. However, if that vendor falls behind the technology or price-performance curves, or stops producing a particular solution, your computer environment becomes obsolete. That's why open systems usually are expected to be multivendor systems.

If open systems are multivendor systems, how do these vendors agree on the way their systems are to interface with each other? If one company (or group of companies) creates or maintains the definition of the interface, it could have a permanent advantage in time and performance over any others who use the interface. Because such a specification is not defined through an open process, it is a proprietary specification, even if implemented by multiple vendors. The first requirement for open systems, then, is that they be based on open standards; open standards are standards developed with an open, consensus-based process, as are all IEEE standards. To paraphrase Woodrow Wilson, open systems require "open standards, openly arrived at."

Interface v. implementation

Interface standards, as opposed to implementation standards, are the second requirement for open systems. An interface is like a set of acceptable building practices for a house. Building practices tell generally how houses should be built, and what kinds of materials should be used for a particular purpose. When an architect designs a house, implementation-specific decisions are made about how many floors the house will have and how many bedrooms are to be built; each decision ultimately will choose either to implement or not to implement building practices. An implementation standard is a plan for a particular house—every house built to an implementation standard would look the same, allowing minimal innovation.

Note that I am not talking about the building codes, which are local and county regulations. These regulatory standards cover things such as safety and sanitation; you must plan to follow them or the county will not issue a building permit. Rather, I am discussing interface standards that make recommendations (36-inch exterior door, 30-inch counter height, 1/2-inch copper plastic pipe). This sort of interface standard facilitates innovation: Walk around and see how many different designs (implementations) can be built that implement the same interface (building practices). Because these are interface standards, and because multiple vendors implement these standards, the average home buyer has a wide choice of standard-size doors, each of which is or can be individualized. If the buyer wants a nonstandard implementation, it can be designed, but the interface may be violated. There will be an associated cost with this variation, including challenges with trying to move appliances (which are built to fit through standardized doors).

Unobstructed access

The third requirement for open systems is that use of the interfaces be free of unreasonable legal, financial, or other restrictions. I mentioned earlier that competition is a critical factor in open systems; real competition isn't practical without free access to the interfaces. Even apparently moderate royalties or innocuous-seeming administrative requirements can stifle competition, to the point that the interface isn't truly open. To use the building example, if there were a $100 fee per door charged to the manufacturers of 36-inch exterior steel doors (for use of the interface called the "36-inch door interface"), the use of the 36-inch door would be very limited. It would be economically unworkable. Similarly, interfaces that allow implementations in the information technology industry must be open—something that is guaranteed by an American National Standard, but not guaranteed by an "industry standard," which is usually a de facto marketing-based activity.

Quality standards

Quality standards are the fourth requirement. In this context, quality refers to the attributes of the interface standard—the interface must be characterized by adequate (though not necessarily maximum) performance, completeness, lack of ambiguity, and conciseness. While meeting these requirements is possible, it requires knowledge and hard work. See my column in last December's issue of *IEEE Micro* for a discussion of quality standards.

The title of this column is taken from the opening scene in Macbeth, in which the three witches gather to make their baneful brew that signals the doom of Macbeth. It is the situation in which the industry now finds itself—fair does seem foul (formal standards are too slow, too complicated and awkward, too rule bound). At the same time, claims for industry standards have become all the rage—but more and more they are proving to be major sources of confusion. Over time, and probably after a certain amount of tragedy, good will triumph, and the benefits of truly open systems and standards will become available to both users and vendors.

A guardedly cheerful note—for a change

Richard H. Stern

Oblon, Spivak, McClelland, Maier & Neustadt, P.C.

1755 Jefferson Davis Highway, Suite 400

Arlington, VA 22202

Over the last year, things have begun to look up a little in copyright law. (I put aside the recent law for "felonizing" certain software professionals, an issue discussed elsewhere in IEEE publications.) In a series of decisions by appellate courts in various parts of the United States, a trend appears to be emerging against "lookie-feelie" and legal metaphysics treatment of copyrights in computer programs. Courts are beginning to substitute common sense for the mumbo jumbo of "sequence, structure, and organization" and "nonliteral aspects" of computer programs.

The past was prolog

A year or two ago, and through most of the 1980s, the trend of court decisions was that computer programs and their copyrights emanated some ineffable miasma that defied any explicit description. But if competitors came too close to whatever it was in developing a competitive software product, they would be held guilty of copyright infringement and heavily mulcted for their insolence. Underfunded software start-up companies were unable to withstand litigation assaults based on these legal doctrines, and repeatedly were compelled to withdraw from the market when sued, or threatened with suit, by established software marketers. Stemming largely from the decision in *Whelan Associates, Inc.* v. *Jaslow Dental Laboratory, Inc.* in 1987, this trend of decision led to suits against competitors based on their copying such expressive elements of plaintiffs' computer programs as the following:

- placing screen captions at the top center of the screen;
- using the color blue as screen background;
- designating which keystrokes a user should press to enter the program function that a given screen menu word designated, by capitalizing and highlighting (making brighter) the letters of the menu word corresponding to the keystrokes;
- labeling the opening menu of a program as "Opening Menu;"
- using pull-down menu windows in reverse video;
- using the same command language to operate program functions;
- using the same commands and keystrokes for given program functions that the plaintiff's earlier program used for those functions;
- having the same list of commands and tasks to be performed;
- using the same switch patterns on a machine's front panel to actuate the machine's software; and
- imitating the plaintiff CADAM's computer program by being "too CADAMish."

In this last item, CADAM, a major CAD/CAM software developer, sued start-ups Adra and Adage for marketing computer programs that copied the "look and feel" of the CADAM program. In addition to charging the defendants with marketing and promoting a "CADAMish" program, the plaintiff complained of the defendants' marketing their program as "CADAM-compatible" and "a CADAM look-alike." (See *IEEE Micro*, Apr. 1986, pp. 64-65.) The defendants apparently exited the market rather than bear the expense of resisting the copyright infringement action.

Mesmerized by analogies that ingenious counsel drew between computer programs and poems, novels, and plays, some courts resolved to protect what they imagined to be the "plot,"

0272-1732/93/0200-0073$03.00 © 1993 IEEE

"style," and "characterization" of computer programs. They anomalously treated copyrights on computer programs as if they were patents.

(The leading US precedent against doing so is *Baker* v. *Selden* in 1879. As the Supreme Court explained in that decision, treating a copyright as if it were a patent defrauds the public, because a patent monopoly is foisted on the public without the built-in protections of the patent system.)

The courts drew the line between unprotected *idea* and protected *expression* at such a high level of abstraction that virtually any competing computer program would be found to have taken *expression* and thus have infringed the copyright. At the same time, they consciously elevated the legal metaphysics of copyright law above the parties' mere "commercial and competitive objectives."

Many of those in the software industry (and probably the overwhelming majority of working software professionals) became convinced that courts were incapable of resolving software rights disputes sensibly. They felt this way because the courts' legal tools were inadequate to the task and because the judges (coming from the wrong one of C.P. Snow's two cultures) could not understand software. As one court recently observed, responsive proposals were to substitute a *sui generis* (unique) software law or "industrial copyright" type of industrial property law for the present law of software copyrights, and to establish an expert software tribunal in place of courts.

Things seemed to have reached a new low point by early 1992. One district court in Massachusetts simply dismissed out of hand the legal relevance of problems in having to learn new and unfamiliar computer program user interfaces (*Lotus Development Corp.* v. *Paperback Software Int'l*) and another district court in San Francisco found disassembly of code unlawful per se (automatically) under the copyright laws (*Sega Enterprises, Ltd.* v. *Accolade, Inc.*, later reversed on appeal).

What's new

Very recently, however, a seemingly contrary judicial consensus has emerged. Quite suddenly, a majority of US courts have rejected the *Whelan* rationale and have said that a copyright on a computer program is *not* a patent, and must be interpreted more modestly. Recognizing the flawed logic of *Whelan* and its progeny, the US Court of Appeals for the Second Circuit (New York) pithily summed up the current thinking in *Computer Associates International, Inc.* v. *Altai, Inc.*, 1992:

> Rightly, the district court found *Whelan*'s rationale suspect because it is so closely tied to what can now be seen with the passage of time as the opinion's somewhat outdated appreciation of computer science. *Whelan*'s approach relies too heavily on metaphysical distinctions and does not place enough emphasis on practical considerations.

Under recent decisions —the Second Circuit's decision in *Altai*, and the Ninth Circuit (San Francisco) decision in *Brown Bag Software* v. *Symantec Corp.*, 1992—a new method of legal analysis for software copyrights has emerged. First, the court filters out all unprotected subject matter (elements dictated by efficiency or external factors, and public domain subject matter) to derive the copyright owner's protected residuum (what is left after subtracting the unprotected subject matter). The court then compares the residuum with the accused work of the defendant. Only if what the defendant took from that residuum (disregarding the rest) was substantial is the defendant is an infringer.

These decisions also recognize the appropriateness of trial courts having their own expert software witnesses assist them in addressing the intricacies of programming's technical issues. Other recent decisions—*Sega Enterprises Ltd.* v. *Accolade, Inc.* in the Ninth Circuit and *Atari Games Corp.* v. *Nintendo of America, Inc.*—establish the legitimacy of disassembly and reverse engineering of computer programs when necessary for legitimate commercial objectives. Somehow, something suddenly became different.

Now what?

Is everything in computer software copyright law now wonderful? Is there no longer any need to fix the system, since at the moment it does not appear to be broken?

Unfortunately, the system may still be badly bent, even if it is not completely broken. The structural problems that led to the many complaints by software professionals and others in the industry remain. That the courts are beginning to learn how to be more rational in applying copyright principles to computer software does not mean that copyright law is a legal scalpel, after all, rather than a blunt instrument. Both the Second Circuit in *Altai* and the Ninth Circuit in *Accolade* warned against "forcing a square peg into a round hole." They meant that when one tries to apply ordinary principles of copyright law to computer software, one gets very peculiar results—sometimes quite startling or bad ones.

Unless we devise a round peg for a round hole (or square off the hole, if you prefer), we shall continue to lurch from one software law crisis to another. That the present crisis seems to have passed is no proper cause for self-congratulation. Future software crises must be anticipated until the structure of software law is mended.

The European Community's *sui generis* database directive, the 1984 US *sui generis* chip topography law (emulated by chip topography laws of many other nations), the Japanese *sui generis* software law proposals of the early

1980s, the WIPO (UN World Intellectual Property Organization) *sui generis* software proposal of the late 1970s, and (catch this) IBM's *sui generis* software proposals around 1970 have all pointed to the right way. We need a properly thought-out *sui generis* utility-model type of law for computer software. It should treat software (at least in its noncode, nonliteral aspects) as the industrial property that it is, not as a species of poem or oil painting. Bridging the two cultures may be a noble idea, but the software industry would experience much less wear and tear if the experiment were carried out at some other experimental subject's expense.

That is not to say that we need software patents as the solution. The three decades of the Algorithm War in the US have shown that patents do not work properly, either, for abstract aspects of software. We need a system that borrows appropriately from copyright law, patent law, utility-model law, and perhaps European imitation law as well. It should combine selected features of each, and new features where the nature of software dictates it, to provide a form of legal protection that properly fits the subject matter to the commercial needs of industry, software professionals, and software users, and to the interests of the public. The task of crafting such a system is not easy or fast, but the alternative is perennial ineptitude and recurrent crisis.

Software Report

David K. Kahaner

US Office of Naval Research, Far East

kahaner@xroads.cc.u-tokyo.ac.jp

Malaysia and Singapore

Malaysia confounded my expectations on my first visit there. From my readings, I had expected to find a Third World country, agriculture driven and visibly regulated by a strong Islamic fundamentalism. Not so, at least from what I could see.

Kuala Lumpur, the capital, looks strikingly like its southern neighbor, Singapore, must have looked only a few years ago. Large residential, commercial, and hotel construction abounds. Ultramodern skyscrapers jostle for space with ancient mosques. Many of the structures are weary, true, but many new department stores and slick malls have gone up as well. Coming soon to a downtown site being vacated by a racetrack is the tallest building in Southeast Asia, over 90 stories tall. A telecommunications tower, being built jointly with the Germans for $100 million, will be over 420 meters high, the world's third tallest tower and Southeast Asia's highest.

Pedestrians and vehicles throng the city's streets and shops: Unlike Singapore, Kuala Lampur has no subway system, so buses are packed. Though many Malaysian women still cover their faces with traditional black Moslem garments, many more wear brightly colored clothing. Western jeans, pants, and tee shirts are everywhere. Stores overflow with the usual cornucopia of Japanese electronics, plus clothing from famous houses around the world. The fashion conscious, sipping their cappuccino and Perrier, crowd the city's cafes.

The road between the capital and the airport (about 30 km) is lined with multinational factories. According to my taxi driver, the downtown Hilton is busy, but not nearly so as the one near the airport, a more convenient stop for international business people visiting their Malaysian subsidiaries. A new international airport, Southeast Asia's largest, will be built about 40 km from the capital at an estimated cost of $8 billion; the old airport will service domestic flights. The road south from Kuala Lumpur toward Singapore is new, multilaned, and spacious, though not completed all the way to the border. The countryside shows substantial evidence of new building, along with plenty of examples of agriculture, primarily rubber and coconut palm plantations. The government has earmarked about $40 billion for infrastructure, social development, and defense expenditure over the next five years. In most areas of economic development Malaysia leads Thailand, and per capita income is almost twice as high.

Malaysia, formerly British-ruled Malaya, gained its independence in 1957, and now is ruled by a constitutional monarch elected on a rotating five-year term basis by the nine hereditary sultans of the traditional Malay states from among themselves. The country occupies the southern half of the Malay Peninsula, which connects through Thailand to mainland Asia, and about half the large island of Borneo to the east. Malaysia has almost 18 million people of whom about 30 percent are of Chinese

0272-1732/93/0200-0075$03.00 © 1993 IEEE

extraction, 9 percent from India or Ceylon (mostly Hindus), and most of the others Malay; almost all the latter are Moslems.

Large and not heavily populated by regional standards, Malaysia is well endowed with natural resources, including lumber, oil, and natural gas. Previously, the British focused on tin and rubber as well as shipping; tin exports now run at a rate of about $300 million, about one fifth the amount obtained from palm oil. The city of Malacca (150 km south of Kuala Lumpur) on the mainland's west coast was Portuguese, then Dutch, then British, and is at the juncture of trade routes between Europe and the Middle East. The adjacent Straits of Malacca are still among the world's busiest waterways.

Recent growth has been very strong, averaging about 8 percent annually since 1980. Unemployment is just over 4 percent, considered full employment, but a shortfall of more than half a million workers is predicted by the end of the decade. Rapid growth has generated a modest amount of inflation (around 4 percent), and the country has a weak balance-of-payments position, the latter fueled by increases in consumer spending and foreign investment. The manufacturing sector claims that its labor pool is already short by 80,000 workers. Many foreign workers, including more than half a million from Indonesia, are employed illegally. At the same time, higher salaries and opportunities elsewhere are attracting skilled Malaysians to move out of the country, a situation Korea, Taiwan, Hong Kong, and other rapidly developing countries in the region have also faced. However, many of these Malaysians are returning to their homeland in senior positions, now that the economic outlook is brighter.

Many Western companies have found a home in Malaysia, and investment from outside Malaysia is very strong, more than $6.5 billion in 1990, with France and Australia involved in two large refinery projects. Taiwan has been Malaysia's largest investor, with almost $5 billion since 1987, though the rate has been reduced recently, as Taiwan has shifted its attention to mainland China and because a $3 billion steel plant project is still on hold. While I was there, Motorola celebrated its 20th anniversary in Malaysia, having invested more than $350 million, and its Malaysian subsidiary has been given the task of spearheading the entry into China. Motorola records substantially more than $1 billion in turnover at four manufacturing facilities here, between 20 and 30 percent of the company's global output.

If current plans are implemented, Malaysia will spend a great deal of money developing its research and development base. By the turn of the century, the country plans to spend 2 percent of its GDP on R&D expenses (1.5 percent by 1995). Most of this increase should come from the private sector whose contribution is predicted to increase to about 60 percent of total expenditures. Five priority sectors have been identified: biotechnology, automatic manufacturing, advanced materials, electronics, and information processing. The current budget allocates about $250 million to strengthen existing R&D institutions and promote joint research between private, university, and government institutes.

SEARCC 92

The 11th annual South East Asia Regional Computer Conference, held this year in Kuala Lampur, was attended by about 650 delegates. Composed of computer professionals from Pakistan, India, Sri Lanka, Thailand, Malaysia, Singapore, Indonesia, Hong Kong, Philippines, Australia, and New Zealand, SEARCC is designed so that information technology (IT) professionals can meet and share information. SEARCC is not primarily a research conference on computer science, although some research activities are featured. This year's conference theme was "IT: Building Information Infrastructure for National/Regional Growth."

At the conference, we learned that Malaysia has officially embraced open systems for public sector procurements, meaning that government agencies that are planning to purchase computer systems, software, and so forth, can specify their requirements in terms of various IEEE, ANSI, and ISO standards for general principles, operating system interfaces, programming languages, commands, utilities, networks, device interfaces, data management, interchange and compression, databases, user interfaces, and security and system development methodology. They can then expect that vendors will be able to comply on the basis of satisfying the standards detailed in these documents. At the moment, agency participation is voluntary. Nevertheless, this is really quite a different situation from say, Japan, where open systems have not been as healthy as their proponents would like.

The conference included much discussion of the status of software versus hardware in Southeast Asia. Most emphatic on this topic was Stan Shih, founder and chair of Acer, Taiwan's largest computer company (more than $1 billion in sales in 1991), and the most respected Asian computer maker outside Japan. Shih recommends moving away from hardware and into software. For the past 10 years developing Asian countries have concentrated heavily on the development of PC-related hardware; this part of the world is now one of the world's leading PC hardware manufacturing centers. But intense competition among PC hardware manufacturers will reduce profit margins, and the future lies in the development of value-added software, primarily in an open system environment. Shih detailed specific steps:

- Develop highly focused and niche products initially, such as firmware bundled products, concentrating on the regional markets in Asia and use PC marketing channels already operational for exporting software.
- Cultivate software experts by training more people. Enlist government support in training personnel

from academic or industrial sources in the development of highly specialized products. Establish software development centers in countries with existing software manpower.
- Attract well-known software houses for local investment by offering incentives. Transferring development technology from these companies would push software produced in Asia forward to world-class standards sooner than by producing software independently.

"Most important," he says, "is the formulation of long-term development strategies, creative and customer-driven marketing, product quality improvement, strong product support, and continuous product research and development that will make a world-class competitor." In my opinion, this kind of philosophy has no relation to what one normally associates with Asian software; if implemented, watch out Microsoft!

Exhibits

More than 50 organizations were represented at the heavily attended exposition that accompanied SEARCC 92. These were mostly vendors demonstrating open system applications and PC/WS commercial hardware products. The PC clone business is slow, and several vendors were offering "fire-sale" prices for 386 and 486 systems, even throwing in computer tables or other encouragements.

One particularly interesting exhibit involved the work at the Center of the International Cooperation for Computerization. CICC, a nonprofit organization founded about 10 years ago by the Ministry of International Trade and Industry (MITI) of Japan, is designed to implement cooperative activities that promote computerization in developing countries. More than 50 Japanese companies participate, and there are activities in almost 20 countries.

CICC's main cooperative research activity is a machine translation system for Asian languages (currently Chinese, Thai, Indonesian, Malaysian, and Japanese), work that has been in progress since 1987 and will run through 1993. In Japan it involves researchers at the Electrotechnical Laboratory (ETL), CICC's Machine Translation System Laboratory, the Japan Electronic Dictionary Research Institute, and various computer manufacturers and software houses. Each of the four other countries also has a research institute associated with the project. CICC has contributed over $3 million toward the project. The main approach is to preedit text to make it easier to translate, followed by morphological, syntactic, and semantic analysis, and eventually conversion into interlingua using the rules of sentence analysis grammar. In other words, an intermediate language is used as the pivot for translation, after which sentences are generated in the target language. Main applications are to translate technical documents at high speed.

Singapore as role model

Meanwhile, Malaysia is trying to copy those aspects of Singapore's development that seem appropriate. No doubt, little Singapore has been a tremendous success, and is an inspiration to it neighbors. Even during the current recession its economy has expanded at a real rate of 5 percent during the first half of 1992, and unemployment is 2 percent. Inflation since 1974 has averaged less than 4 percent (US average during this same period was about 6.5 percent), and this year it should be roughly 2.5 percent, about one third of the average wage increase. Singapore's 1991 per capita GDP was $20,400, compared to $14,900 in 1984 (this corresponds to a GNP of $13,271 in 1991). The future also looks very bright. Economists have predicted that Singapore is very likely to be among the 20 richest countries in the 21st century. To do that it has to continue to focus on people and seven major industries: microelectronics, biotechnology, new materials, civilian aviation, telecommunications, robots and machine tools, and computers and software. Success will come if other countries in the area allow Singapore to become the headquarters city for the region, while they are also moderately successful themselves.

Singapore's government has a very definite slant to economic development. "It is Singapore versus other countries," says Singapore's Prime Minister Chok Tong Goh as he places Singapore's team approach squarely between Hong Kong's every man for himself and New Zealand's state welfare approach. (Goh singles out New Zealand as a case of what not to do; a country that was fifth richest in 1966 and is now 19th, while Singapore has gone from 33rd to 18th during that same period. Goh's explanation: New Zealand's ranking fell because its welfare subsidies increased the dependency of the people and sapped their competitive drive.) According to Goh, the key is giving people incentives to strive: good pay and light taxes. (Singapore's beginning tax rate is 3 percent, compared to 15 percent and 30 percent in Japan and Sweden; half of Singapore's taxpayers, about 500,000, pay $100 or less in taxes.)

Goh also wants to make Singaporeans asset owning. Currently, only 14 percent of adults own shares in publicly listed companies (compared with 21 percent in the UK and 27 percent in Japan), and Goh hopes to increase that to 30 percent. The government plans to sell shares in Singapore Telecom at a discount next year, and also plans to sell shares in the Mass Rapid Transit, Port of Singapore, and a new company formed to run the country's electricity and gas departments.

Good books and good software

Richard Mateosian

2919 Forest Avenue

Berkeley, CA

94705-1310

(510) 540-7745

I look at many books and software packages in the course of preparing this column. I select items for review that I think you will find interesting or important in your work. I also try to select products that are worth your trouble and expense in obtaining and using them. In other words, I like to choose products that I can recommend enthusiastically. Negative reviews have an important role to play in other contexts, but I think that positive reviews are more useful here.

Since there are so many more good products than I can provide in-depth reviews of, I've decided to give you a potpourri of short reviews this month. Please let me know if you find this approach useful. If so, I'll do it again from time to time.

Books

Debugging—Creative Techniques and Tools for Software Repair, Martin Stitt (Wiley, New York, 1992, 432 pp.; $32.95)

This book is a gold mine. It deals almost exclusively with assembly-level debugging, principally for the Intel 80x86 architecture running the MS-DOS operating system. What Stitt says within that framework comes from his obviously deep understanding of more general principles.

Stitt wants to teach you how to approach software performance anomalies. He wants you to forget stereotypes about "the black art of debugging." He wants you to adopt a disciplined, systematic approach to problems. This approach requires you to diagnose with cool detachment, attend carefully to detail, and never lose sight of the forest for the trees. Debugging may not be your favorite activity, but as Stitt points out, the better you are at it, the less time you'll have to spend doing it.

I've seen many books that attempt to teach disciplined, systematic approaches to software tasks, and most of them aren't worth the paper they're printed on. This one is different. Stitt demonstrates in his writing the same detatched analysis, attention to detail, and broad view that he wants you to adopt in debugging.

I've been programming computers since 1960. I've always enjoyed and had great success at debugging my own programs and those of others. This is the first decent account I've seen of the problems and techniques of debugging. I began my evaluation of this book by opening it at random to about a dozen different places and reading a paragraph or so at each place. Each time my reaction was "yes, yes, yes." Assembly-level debugging may not appeal to you, but if you do any programming at all, you can probably benefit from this book.

Inside Windows NT, Helen Custer (Microsoft Press, Redmond, Wash., 1992, 416 pp.; $24.95)

David Cutler, who led the designs of Digital Equipment Corp.'s RSX-11M and VMS operating systems, came to Microsoft in October 1988 to lead the development of their next-generation operating system. Windows NT is the result. When it's finally ready—probably some time this year—it will take its place at the high end of the Microsoft line, providing upward compatiblity for DOS and Windows applications.

Helen Custer spent three years as part of the Windows NT design team. Her job was to write this book. Before starting, she read Tracy Kidder's *The Soul of a New Machine* for inspiration, but her book is not meant to be anything like Kidder's. Custer's book focuses more on the structure of the final product than on the human and intellectual story of its creation. She mentions

0272-1732/93/0200-0078$03.00 © 1993 IEEE

the names of members of the design team and gives them credit for their specific contributions, but that's as far as she goes with the human element.

Custer starts from market needs and design goals, gives an overview of the resulting design, then spends the rest of the book giving an in-depth view of the operating system components. Depending on how much you care about such things, you will find this material somewhere between deadly dull and intensely interesting. Wherever you fall on that spectrum, you'll probably appreciate Custer's clear writing style and the book's open format. Custer has written an accessible account of an important new system.

If you want to understand Windows NT, this is the authoritative account. It will probably be the best book on the subject for a long time to come.

The Elements of Friendly Software Design—The New Edition, Paul Heckel (Sybex, Alameda, Calif., 1991, 349 pp.; $22.95)

The original edition of this book appeared in 1984. The new edition contains the original edition as an unmodified subset. The new material tells the story of Heckel's battle to assert his patent rights against the giants of the computer industry, particularly IBM. Heckel presents his side persuasively and generalizes to the problems faced by all inventors, but it is still only his side of the story. I found it fascinating, but it is of much less general interest and importance than the original material.

Paul Heckel is an original thinker. His fundamental message is that software design is a form of communication. This metaphor allows him to draw immediate parallels between software design and other forms of communication, notably film. This thought process leads him to 30 maxims, which he expands upon with examples from software design situations and from everyday life.

Heckel tells us that we have to overcome our instincts before we can design friendly software. These counterproductive instincts are:

> We think logically, not visually.
> We base our designs on our knowledge, not the user's.
> Our programs evaluate our user's actions.
> We make our programs take control.
> We think in generalities, not specifics.
> We structure for internal organization.
> We strive for a program's internal simplicity.
> Our knowledge constrains our vision.

In a newly added chapter written with Chuck Clanton, Heckel says that the most critical aspect of user interface design is the design of conceptual models. These facilitate communication between the designer and the user and form a framework that the user can become comfortable in. The most helpful conceptual designs are metaphors, that is, analogies with real-world situations. These allow the user to bring existing skills and knowledge into the new situation.

Heckel moves from theorizing about metaphors into describing his own card-and-rack metaphor. He compares and contrasts it with the well-known desktop and spreadsheet metaphors. This is interesting material, but ties again into his patent problems.

At one point Heckel quotes Blaise Pascal, "Anything that is written to please the author is worthless." I hope Heckel will take that message to heart and will someday bring out a version of the book that finds a better way to communicate the lessons of his recent problems. Very little in this fine book can be considered worthless. But there is a distinct difference in perspective between the parts that teach friendly software design and the parts that document and support his business struggles.

Until that new version comes out, you should buy this one. It's still the best book on user interface design.

µC/OS—The Real-Time Kernel, Jean J. Labrosse (R&D Publications, Lawrence, Kansas, 1992, 284 pp.; $29.95)

This is an extremely instructive book. It's not a polished job of publishing, and the text could use professional editing, but the subject redeems all of that.

Real-time kernels are important in embedded systems, but few books have been written about them. Companies like Ready Systems and Wind River have developed excellent products in this area, but they are not in a hurry to give away their secrets.

Labrosse understands the requirements, many of them counterintuitive, of real-time systems. He has written a real-time kernel in C with a small amount of carefully isolated assembly language. His book is essentially an annotated listing of that kernel. A separately available diskette contains the entire source code.

Obviously, this kind of book is not for everyone. For the person who works with embedded systems, this book is worth looking for.

Software

Microsoft Word 5.1 for the Macintosh and Word for Windows 2.0 (Microsoft Corp, Redmond, Wash.)

I've been using Microsoft Word for the Macintosh for a long time—on my original Macintosh, on its successor the Mac Plus, and on my current SE/30. It's a powerful, full-featured word processor, and I like it very much. Until now, it has always been better than the corresponding product for the PC. Now, however, Word for Windows is at least as good as Word for the Macintosh. In some ways it's much better.

Of course, there are the differences in the platforms. My Macintosh SE/30 has a tiny black-and-white screen, while my PC has a super VGA color display of more than twice the area. My SE/30 has a 16-MHz 68030 processor, a 40-Mbyte hard disk, and 4 of its 8 Mbytes

of memory are sitting in a drawer, waiting to be reinstalled. My PC has a 33-MHz 80486, a 200-Mbyte hard disk, and 16 Mbytes of main memory.

It's depressing to have less than 60 percent of my screen available for text when I run Word 5.1 for the Macintosh. It's depressing when my screen saver runs a banner across the top of my screen saying that it doesn't have enough memory to run. But beyond these psychological effects, Word for Windows excels Word for the Macintosh in its conceptual model.

The features of Word for Windows are organized around styles, document templates, fields, and the Word Basic macro language. Word for the Macintosh has no macro facility and uses ad hoc approaches to indexing and other applications of fields. It seems to be evolving toward document templates. Both versions handle styles similarly.

The Windows operating system has an object linking and embedding (OLE) feature, which allows dynamic linkage between files. Apple's System 7 operating system for the Macintosh has a similar capability. Word for Windows seems to make better use of this kind of file linking than Word for the Macintosh does.

Word will probably remain my word processor of choice in the future, but I may take the plunge and move from the Macintosh to Word for Windows.

MKS Toolkit 4.1 for DOS (Mortice Kern Systems, Waterloo, Ontario, Canada; US$299)

If you're used to Unix and you have to use DOS, this package can give you all the comforts of home. The package contains a complete implementation of the Korn Shell, uucp, the vi editor, an excellent implementation of awk, a make facility, pipes, tar, and all of the most popular Unix utilities. All told, the package gives you a 3-inch stack of manuals and about 6 Mbytes of programs, examples, and on-line tutorials and documentation.

The relatively painless installation procedure also sets up a rudimentary Windows interface to some of the tools. This looks nice but doesn't really add much to the basic tool set, since Unix tools are all essentially optimized for use from the command line.

This package is designed for programmers, but anyone familiar with the Unix environment will appreciate it immediately. If you use DOS and you don't know much about Unix, this is a good way to find out what all the fuss is about. Be careful—you might not be able to go back to DOS.

Speed Reader Windows Version (Davidson & Associates, Torrance, Calif.; $49.95)

This is a straightforward training package to improve your reading skills. There are no gimmicks. The authors have incorporated well-known principles of reading into a neat package. They have integrated the package competently, if not elegantly, into the Windows environment.

There are six basic activities: warm-ups, eye movement, newspaper reading, paced reading, timed reading, and the Eye Max peripheral vision exercise. The program lets you log in by name and keeps track of your progress on the various activities. You can examine a log of your sessions or look at bar graphs of your progress. The package keeps track of your reading speed and comprehension level for each type of activity.

If you've ever played computer games and watched your scores rise as your competence improved, here's a chance to try the same process to develop a useful skill.

Reader Interest Survey

Indicate your interest in this department by circling the appropriate number on the Reader Service Card.

Low 183 Medium 184 High 185

Microelectronic Systems Branch at Goddard. "The silicon alternatives we evaluated were either too slow or consumed too much power," he added.

The GaAs chip forms part of the second Tracking and Data Relay Satellite ground station upgrade at White Sands, New Mexico, and could support the future deployment of Space Station Freedom and the Earth Observing System. It features a programmable search, check, and lock strategy for synchronization of data frames up to 32 Kbits in length and provides double-buffering of output data. Standard microprocessor control logic using standard TTL control signals controls the device.

Vitesse Semiconductor Corporation headquarters in Camarillo, California.

Editorial Board changes

Editor-in-Chief Dante Del Corso announces several changes in *IEEE Micro*'s Editorial Board. Board member Maurice Yunik will join K.E. Grosspietch and Ashis Khan as Associate Editors in Chief. Yunik, of the University of Manitoba, will support Del Corso in seeking and reviewing manuscripts from US and Canadian authors.

Del Corso also welcomed three new Board members: Stephen L. Diamond, Osamu Tomisawa, and Uri Weiser.

Diamond is director of standards at SunSoft, Inc., in Mountain View, California. He is chair of the IEEE Microprocessor Standards Committee, Policies and Procedures, chair of the Computer Society Standards Activities Board, and a member of the US delegation to ISO/IEC JTC1 SC 26, the Posix Executive Committee, and the X/Open and Sparc International boards and committees. He will reprise the magazine's Micro Standards column (see p. 71 this issue).

Tomisawa and Weiser will speed the review of manuscripts for *Micro*.

Tomisawa manages the Microcomputer Department B at the Kita-Itami Works of Mitsubishi Electric Corporation, where he works on memory and logic VLSI design. He is a member of the IEEE and the Institute of Electronics, Information, and Communication Engineers of Japan, and an associate editor of *IEICE Transactions on Electronics*.

Weiser is Microprocessor Group manager, Platform Architecture Center, Microprocessor Architecture Development for, Intel Israel in Haifa. He has served as chair and a member of the Program Committee for a variety of conferences and symposiums including ICCD, Computer Architecture, Hot Chips IV, and CompEuro.

Literature

Technology trends, key issues, opportunities, and market growth rates form the major part of this study on the RISC market. *"RISC Impact on the Computer and Workstation Markets," Electronic Trend Publications, Saratoga, CA; (800) 726-6858, ext. 1091; $495.*

Database programers at any level who plan to develop applications for the Clipper 5.0 should benefit from this 1,351-page book by Joseph D. Booth. It includes an introduction to the basics and advanced networking, debugging, and pop-up programming information. *Clipper 5: A Developer's Guide, M&T Books, San Mateo, CA; (800) 688-3987; $44.95, book and disk.*

Micro bits

• Vitesse Semiconductor is sponsoring a **design contest** that illustrates the use of the Viper GaAs gate array. Entries must be postmarked by March 31, 1993; prizes will be awarded. If interested, call Vitesse Marketing at (805) 388-7455.

•Striving for a tenfold increase in **semiconductor manufacturing productivity** is Texas Instruments' Microelectronics Manufacturing Science and Technology project. Funded by DARPA and USAF Wright Laboratories, MMST uses object-oriented programming and database technology and "revolutionary concepts."

•**TI and IDT** signed an alternate source agreement for logic devices with built-in boundary scan. Each will offer advanced bus interface and LSI controllers that comply with the JTAG/IEEE 1149.1-1990 testability specifications.

•**Wireless LANs** may capture 17 percent of all LAN shipments by 1997, according to BIS Strategic Decisions, Norwell, Mass. The reason: improved economics from wired networks, standards activity, and the emergence of more mobile computing devices.

•The Dataquest market research firm lists Motorola as the leading worldwide supplier of **8-bit microcontrollers**, ranking the 68HC05 and 68HC11 first and ninth in worldwide shipments.

Reader Interest Survey

Indicate your interest in this department by circling the appropriate number on the Reader Service Card.

Low 198 Medium 199 High 200

Send information for inclusions in Micro News one month before cover date to Managing Editor, IEEE Micro, *PO Box 3014, Los Alamitos, CA 90720-1264.*

ISO's smart highway TC

Noting the significant US and overall international community interest, the International Organization for Standardization Technical Board established a new technical committee for intelligent vehicle/highway systems, which it calls Road Transport Informatics. The TC's scope will include standardization in the field of smart highways, Advanced Traveler Information Services, Advanced Traffic Management Systems, Advanced Vehicle Control Systems, Advanced Public Transportation Systems, and Commercial Vehicle Operation. Pending approval by the ISO council, the Technical Board decided to allocate the secretariat for this committee to the US through the American National Standards Institute.

Standardization work for smart highways is also taking place in the European CEN, CENELEC, and ETSI committees; in addition, the International Electrotechnical Commission proposes to establish a new technical committee for road traffic signal systems.

For further information contact ANSI at 11 West 42nd Street, New York, NY 10036.

US to participate in Japan's Real World Computing program

The US and Japanese governments plan a joint prototyping project to further the design and development of advanced computing technologies that combine light-wave and electronic components. The hybird systems to be worked on would serve as a bridge between today's electronic computers and the fully optical, parallel processing machines envisioned for the future.

Part of Japan's 10-year, $500-million Real World Computing program for information processing, the new optoelectonics project will involve researchers and processing facilities in both nations. A 10-member joint management committee with five representatives from each country will guide the project; Judson French, National Institute of Standards and Technology, will chair the US group. Plans call for establishing a service that links designers of optoelectronic devices and modules with production facilities, or "foundries," through a broker. Each country will select its own broker and arrange the funding for its participants. Japan's MITI will finance the broker in both the US and Japan.

Although the collaboration forms only a small component of the overall RWC program, it allows both countries to develop a model for cooperative research that could lead to other cooperative projects.

For more information, contact the White House Office of Science and Technology Project, Old Executive Office Bldg., Room 428, Washington, DC 20500; (202) 456-7710.

NASA picks 15,000-gate GaAs ASIC

The US National Aeronautics and Space Administration's Goddard Space Flight Center received functional prototypes last fall of a 15,000-gate chip for use in telemetry acquisition systems. The Vitesse Semiconductor Telemetry Frame Synchronizer was implemented in the GaAs Fury VSC15K gate array that is manufactured using the company's proprietary H-GaAs process technology. Anticipating superior performance and low power in the ASIC, NASA selected it over competing silicon bipolar and BiCMOS devices. The synchronizer boosts the upper limit of this type of system performance to 300 Mbps.

"Our requirements called for a high-performance ASIC that could integrate a lot of lower complexity ECL devices into one chip. We chose Vitesse's H-GaAs technology, not only because it offered the speed and integration we needed but because it allowed us to use traditional air cooling," said Jim Chesney, NASA's head of the

0272-1732/93/0200-0081$03.00 © 1993 IEEE

Editorial comments

I liked: _____

_____

_____

I disliked: _____

_____

_____

I would like to see: _____

_____

_____

Reviewers Needed. *If interested, send professional data to Dante Del Corso, Dipartimento di Elettronica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129, Torino, Italy.*

For reader service inquiries, see other side.

PLACE POSTAGE HERE

PO Box is for reader service cards only.

IEEE Micro
PO BOX 16508
NORTH HOLLYWOOD CA 91615-6508
USA

Editorial comments

I liked: _____

_____

_____

I disliked: _____

_____

_____

I would like to see: _____

_____

_____

Reviewers Needed. *If interested, send professional data to Dante Del Corso, Dipartimento di Elettronica, Politecnico di Torino, C.so Duca degli Abruzzi, 24, 10129, Torino, Italy.*

For reader service inquiries, see other side.

PLACE POSTAGE HERE

PO Box is for reader service cards only.

IEEE Micro
PO BOX 16508
NORTH HOLLYWOOD CA 91615-6508
USA

NO POSTAGE NECESSARY IF MAILED IN THE UNITED STATES

BUSINESS REPLY MAIL
FIRST CLASS PERMIT NO. 38 LOS ALAMITOS, CA

POSTAGE WILL BE PAID BY ADDRESSEE

IEEE COMPUTER SOCIETY
PO BOX 3014
LOS ALAMITOS CA 90720-9804
USA

READER SERVICE CARD

IEEE MICRO

February 1993 issue (card void after August 1993)

Name _____

Title _____

Company _____

Address _____

City _____ State _____ Zip _____

Country _____ Phone(_____) _____

Reader Interest

(Add comments on the back)

Readers,
Indicate your interest in articles and departments **by circling the appropriate number** (shown on the last page of articles and departments):

150	151	152	180	181	182
153	154	155	183	184	185
156	157	158	186	187	188
159	160	161	189	190	191
162	163	164	192	193	194
165	166	167	195	196	197
168	169	170	198	199	200
171	172	173	201	202	203
174	175	176	204	205	206
177	178	179	207	208	209

Product Information

(Circle the numbers to receive product information)

1	21	41	61	81	101	121	141
2	22	42	62	82	102	122	142
3	23	43	63	83	103	123	143
4	24	44	64	84	104	124	144
5	25	45	65	85	105	125	145
6	26	46	66	86	106	126	146
7	27	47	67	87	107	127	147
8	28	48	68	88	108	128	148
9	29	49	69	89	109	129	149
10	30	50	70	90	110	130	
11	31	51	71	91	111	131	
12	32	52	72	92	112	132	
13	33	53	73	93	113	133	
14	34	54	74	94	114	134	
15	35	55	75	95	115	135	
16	36	56	76	96	116	136	
17	37	57	77	97	117	137	
18	38	58	78	98	118	138	
19	39	59	79	99	119	139	
20	40	60	80	100	120	140	

READER SERVICE CARD

IEEE MICRO

February 1993 issue (card void after August 1993)

Name _____

Title _____

Company _____

Address _____

City _____ State _____ Zip _____

Country _____ Phone(_____) _____

Reader Interest

(Add comments on the back)

Readers,
Indicate your interest in articles and departments **by circling the appropriate number** (shown on the last page of articles and departments):

150	151	152	180	181	182
153	154	155	183	184	185
156	157	158	186	187	188
159	160	161	189	190	191
162	163	164	192	193	194
165	166	167	195	196	197
168	169	170	198	199	200
171	172	173	201	202	203
174	175	176	204	205	206
177	178	179	207	208	209

Product Information

(Circle the numbers to receive product information)

1	21	41	61	81	101	121	141
2	22	42	62	82	102	122	142
3	23	43	63	83	103	123	143
4	24	44	64	84	104	124	144
5	25	45	65	85	105	125	145
6	26	46	66	86	106	126	146
7	27	47	67	87	107	127	147
8	28	48	68	88	108	128	148
9	29	49	69	89	109	129	149
10	30	50	70	90	110	130	
11	31	51	71	91	111	131	
12	32	52	72	92	112	132	
13	33	53	73	93	113	133	
14	34	54	74	94	114	134	
15	35	55	75	95	115	135	
16	36	56	76	96	116	136	
17	37	57	77	97	117	137	
18	38	58	78	98	118	138	
19	39	59	79	99	119	139	
20	40	60	80	100	120	140	

SUBSCRIPTION CARD

SUBSCRIBE TO IEEE MICRO

All the facts about today's chips and systems

❑ **YES**, sign me up!

If you are a member of the Computer Society or any other IEEE society, pay the member rate of only $23 for a year's subscription (six issues).

Society: _____

IEEE membership no: _____

Society members: Subscriptions are annualized. For orders submitted March through August, pay half the full-year rate ($11.50) for three bimonthly issues.

_____
Full Signature Date

Name _____

Street _____

City _____

State/Country ZIP/Postal Code

❑ **YES**, sign me up!

If you are a member of ACM, ACS, BCS, IEE (UK), IEEE but not a member of an IEEE society), IECEJ, IPSJ, NSPE, SCS, or other professional society, pay the sister-society rate of only $42 for a year's subscription (six issues).

Organization: _____ Membership no: _____

For information on airmail option, send fax request to (714) 821-4010.

❑ Payment enclosed *Residents of CA, DC, Canada, and Belgium add applicable sales tax.*

❑ Charge to ❑ Visa ❑ MasterCard ❑ American Express

Charge-card number

_____

Expiration date

____ ____
Month *Year*

Prices valid through 12/31/93

293 MICRO

Charge orders also taken by phone:
(714) 821-8380 8 a.m. to 5 p.m. Pacific time
Circulation Dept.
10662 Los Vaqueros Circ., PO Box 3014
Los Alamitos, CA 90720-1264

New Products

Send announcements of new microcomputer and microprocessor products to Managing Editor, IEEE Micro, PO Box 3014, Los Alamitos, CA 90720-1264.

Joe Hootman

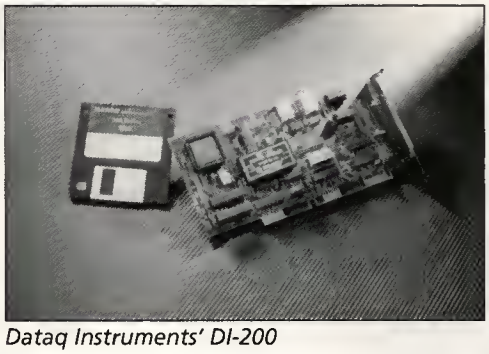
University of North Dakota

DSP software, hardware

I/O card operates in Windows

First to be released in a 200 Series of hardware products is the DI-200 data acquisition card designed to operate in the Windows and DOS programming environments. The 16-channel analog I/O card incorporates DSP-based technology, channel-by-channel programmability, and an 83-kHz burst sampling rate that minimizes channel skew. Promising 12-MIPS performance, the 16-bit DI-200 offers bipolar measurements from ±1.25V to ±10VFS (full scale) or ±10mV to ±10VFS and unipolar measurements from 0 to +1.25V, 0 to 10V (A_v = 1) or 0 to +10mV, 0 to 10V (A_v = 1,000). *Dataq Instruments; $795, delivery from stock.*

Reader Service No. 10



Dataq Instruments' DI-200

Achieve 200-Mflops peak speeds

The MZ 7770 DSP VMEbus module features four interconnected TMS320C40 DSPs for interprocessor communication at 200-Mflops peak speeds. Each C40 with zero-wait-state SRAM holds three more 20-Mbyte/s communications ports that ease interconnections of C40s from multiple boards. Multiple MZ 7770s can be arranged in 3D-mesh, ring, or hypercube multiprocessor architectures. The 6U-size board suits a variety of signal and parallel processing applications and comes with an ANSI-compatible C compiler with a parallel processing runtime library. Additional software includes a C source-level debugger, Texas Instruments' pDSP XDS 510 in-circuit emulator with JTAG diagnostic support and the NOS operating system; an Ada compiler; and the SPOX, Helios, OS-9, and VxWorks operating systems. *Mizar; from $15,900.*

Reader Service No. 11

Real-time VMEbus coprocessor

The 1.1-billion operations/s VMEbus DSP coprocessor called Hydra has added the Helios real-time operating system for development and execution of applications that run on large multiprocessor networks of up to 100 Hydras. Included with Helios are Unix-like PC- and Sun-based cross-development tools plus a real-time multitasking, multithreaded system that runs on the Hyrdra-based target system. The cross-development tools include ANSI C and Fortran compilers, TCP/IP networking, X Windows and Microsoft Windows graphics support, and Posix and BSD libraries.

Helios also supports interprocess communications and synchronization mechanisms including shared-memory locks and semaphores. Programmers can establish communications between multiple programs without specifying the physical connections by making a read or write call to a file descriptor. *Ariel; $3,500 (Helios), from $9,995 (Hydra).*

Reader Service No. 12

TI introduces the C52, enhances C5X products

Promising high performance and low cost, Texas Instruments introduced its latest DSP chip and enhancements for its product line. Company spokesmen say the 16-bit, fixed-point TMS320C52

0272-1732/93/0200-0083$03.00 © 1993 IEEE

DSP for telecommunications and other high-performance applications represents two to four times better performance than the popular C25. The 100-pin thin QFP device performs an instruction in 25, 35, or 50 ns for 40-MIPS execution at either 3.3V or 5V. Designed with a superset of the C25 memory and peripherals, the C52 features a 1K-RAM/4K-ROM configuration, a single serial port, and a single timer.

The C5X 16-bit DSP family also includes the C50, C51, and C53, each with an instruction set that is source-code compatible with C1X and C2X 16-bit DSPs. Enhancements include on-chip power management circuits that provide power consumption in active mode, 2.5-mA/MIPS at 5V and 1.5-mA/MIPS at 3.3V, as well as two power-down modes. *Texas Instruments; $15.95 (1,000s) and $10 (100,000s); C5X volume production 2Q93.*

Reader Service No. 13

Software release supports Windows

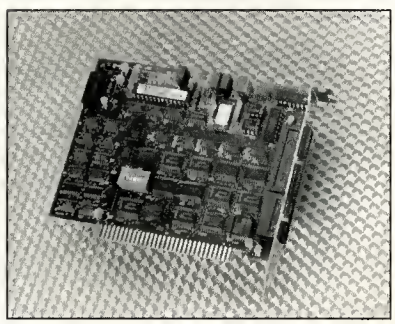
Hypersignal-Windows RT-3 is an integrated signal processing software package of data acquisition, real-time DSP, graphical analysis, visually programmed algorithm development, and DSP development tools, all of which work together. The just-released Version 1.30 features extended snap-in digital filtering, enhanced graphing capabilities such as waveform overlay and 2D and 3D frequency displays, a larger function library with user-written C-compiled blocks, and high-accuracy frequency markers on the spectrum analyzer. According to the manufacturer, the Hypersignal-Windows RT-3 package supports 20 DSP/acquisition boards for real-time instruments. *Hyperception.*

Reader Service No. 14

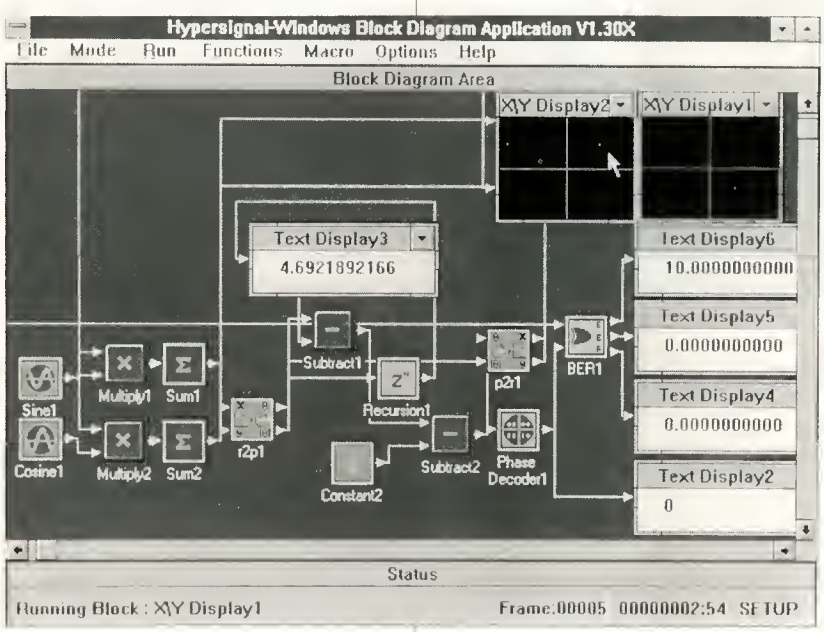
Board consumes 1W power

A 5.3-inch data acquisition board that uses 1 watt of power supports remote and portable applications with 16 single-ended or eight differential analog input channels (12-bit resolution). The PCI-20377W-1 features a 45-kHz throughput rate; programmable gains of 1, 10, 100, and 200; 16 protected digital I/O channels; and a rate generator. A 16-word FIFO buffer ensures continuous data flow to the host when the host is temporarily unavailable. All user-selectable configuration features such as gain, signal range, and single-ended/differential modes are software controlled. The board includes Master Link software libraries for DOS and Windows environments and the Syscheck system assurance utility. *Intelligent Instrumentation; $495.*

Reader Service No. 15



Intelligent Instrumentation's PCI-20377W-1



Hyperception's Hypersignal-Windows

Software

RS/6000 development tools

Now running on IBM RISC System/6000 workstations are Intel's i960 and 8086 microprocessor development tools: the ANSI C cross compiler, macro cross-assembler, and Xray debugger. Xray debugs optimized C code, supports instruction-set simulation, and features the X Windows System Motif interface. The optimizing C compilers comply with the ANSI C standard and accept programs written in the original C language as defined by Kernighan and Ritchie. The C++ compilers comply with Version 2.1 of the AT&T specification. *Microtec Research; from $4,300.*

Reader Service No. 16

Desktop Design Architect

Design Architect PCX lets users run it and Falcon Framework under X Windows while the actual application takes advantage of computer power elsewhere on the network. The package supports schematic entry, remote simu-

lation and synthesis, and QuickSim II in X configurations for graphical display on an X terminal. The company promises to qualify and fully support Sparc workstations with Open Windows 3, HP-PA workstations with HP/UX 9.0, and a variety of other X terminal configurations. *Mentor Graphics; $7,500 per user (3-6 user configurations).*

Reader Service No. 17

Windows I-CASE tool

Version 5.0 of Visible Analyst Workbench I-CASE supports the Microsoft Windows operating environment. The integrated tool set features forward- and reverse-engineering capabilities such as SDM, Yourdon Structured Method, and Information Engineering. Users can generate SQL database schemas, Cobol source code, and C source code from designs developed in the system. Other enhancements include multipage document support, ease-of-use, model navigation improvements, control bar support, repository data access, and text editing. Version 5.0 is upward-compatible with Versions 4.2 and 4.3. *Visible Systems; from $1,895.*

Reader Service No. 18

ASCII, math packages

Turbo Spring-Stat Text Editor II, an ASCII data file editor, supports 64K windows of different files as free memory allows, an optional mouse, and a clipboard; it does not require Windows to operate. Each window with scroll bar is movable and resizable, letting users cut and paste within and between files. Text Editor II requires MS-DOS 2.0 or higher, a 512-Kbyte RAM, and CGA/EGA/VGA or compatible graphics capability.

The Equator II menu-driven mathematical equation storage, evaluation, and plotting system lets IBM users save equations, document variables and parameters used, and evaluate expressions to create tables, graphs, or disk files. Equator II users can also import data files from other sources for plotting. Results may be viewed either on the screen or sent to an Epson or compatible dot matrix printer, an HP LaserJet, or an HPGL plotter. Equator II requires MS-DOS 2.1 or higher, CGA/EGA/VGA graphics capability, 512 Kbytes of RAM, and two 720-Kbyte floppy drives. *Dynacomp; $39.95 (Text Editor II), $79.95 (Equator II); 20-percent discount if order accompanied by this page.*

Reader Service No. 19

Simulation models for PLDs

Behavioral models for the Altera Multiple Array Matrix (MAX) 7000 programmable logic devices have been added to the Logic Modeling Smart Model Library. This 6,500-component library interfaces with MAX+Plus II development tools for accurate modeling of functional and timing delays in an Altera-compiled PLD. Smart Models run on Verilog, QuickSim II, ViewSim, CADAT, HiLo, and VHDL simulators on most Unix workstations. *Altera Corporation and Logic Modeling; shipped with subscriptions/updates (new models), $10,000 per workstation (full library license).*

Reader Service No. 20

Visual Basic gains database manager

Agility/VB lets Visual Basic programmers create database applications using custom controls without writing a line of code. The package includes grid, text, button, and picture controls, and a set of commands that provides program control over database applications. A View Editor tool specifies relationships between multiple, different-format databases in a view so programmers can see them as a single flat file while maintaining all relations and indexes. An Agile Assistant programming aid helps users manage database-related programming tasks.

Agility/VB supports dBase and text file formats and provides its own database for variable-structure and variable-length data storage. The manager requires Microsoft Windows 3.X, Visual Basic 1.X or higher, and an 80286 processor; 2 Mbytes of RAM is recommended. *Apex Software Corporation; $189.*

Reader Service No. 21

LabWindows adds C++ libraries

LabWindows for MS-DOS Version 2.2.1 instrumentation software now includes stand-alone libraries for the Borland C++ and Turbo C++ compilers and Microsoft Visual Basic for DOS (VBDOS) compiler. Version 2.2.1 offers float data type DSP Analysis Library, new cursor functions, DPMI memory manager capabilities, and a library for performing DOS file and directory commands directly from LabWindows.

The C++ libraries let users access the Borland compiler and linker from within LabWindows to create executable programs or add LabWindows libraries to the Borland Interactive Development Environment for program development. Each of the libraries has a Borland-compatible help file that users can load into the IDE for on-line help. Basic programmers in VBDOS can incorporate LabWindows instrumentation functionality into their applications, access the VBDOS compiler from within LabWindows, or load the LabWindows libraries into VBDOS as an external Quick Library. *National Instruments; free upgrades to 2.2 users, $195 for upgrades from previous versions.*

Reader Service No. 22

EDI translator/manager

The Electronic Data Interchange EDI*Transit translation and management system works in both Unix and MS-DOS environments. The program reduces EDI document processing time and features mapping capability, translation of all key standards, task scheduling, and functional acknowledgment tracing. In addition, predefined communication scripts allow easy access to the company's EDI*Express Service. *GE Information Services.*

Reader Service No. 23

Process control enhancements

ExpressLite, a recent release of the Express event management and control system for process control and factory automation applications, offers enhanced graphics, communications options, I/O drivers called Opto-22 Optomux and Modicon V984, and an historical trend-recording feature. Included with ExpressLite is a demonstration application supplied with source code that users can run, modify, or replace with a custom application. ExpressLite supports all Express functions but no actual I/O drivers, up to 256 simulated I/O points, one terminal, and one printer. *Forth, Inc.; $195 (ExpressLite evaluation version), $6,875 (Express)*

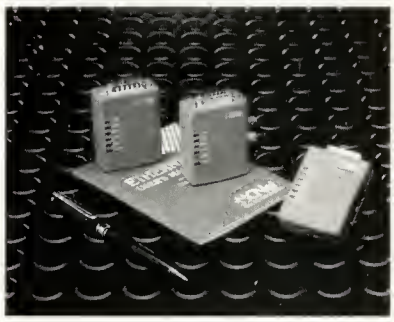
Reader Service No. 24

Communication devices, software

Transceivers fit in vest pockets

The 2-oz. CN815E and CN825E transceivers offer an upgrade path from coaxial cable. The CN815E AUI-to-10BaseT interface and CN825E coax-to-FOIRL (fiber optical interrepeater link) converter support PC, Macintosh, and Sparc Station platforms and are compliant with 10BaseT and fiber optic Ethernet standards. Each 2.28 × 1.79 × 0.9-inch transceiver includes automatic polarity correction and status LEDs for power, transmit, receive, collision, link, and jabber (SQE) signal display. *CNet Technology; $129 (CN815E), $399 (CN825E).*

Reader Service No. 25



CNet Technology's transceivers

Managing tolls and traffic

An IVHS platform system for electronic toll collection reduces traffic congestion, fuel consumption, and auto emissions by allowing motorists to travel nonstop through toll lanes. According to the manufacturer, the New Hampshire State Police tested the patented radio frequency identification technology at speeds in excess of 90 mph.

The microprocessor-based read-write device improves on the read-only process that uses either barcode tags or radio-reflective tags to read a passing vehicle's ID. Read-write allows information, such as the entry point of a turnpike for later toll calculation, to be written onto an intelligent transponder placed in a vehicle. Like a postage meter, the transponder is electronically charged with a value, and that value is reduced each time the car passes through a toll lane. An LCD display and audio alarm on the device give the motorist real-time infomation on the remaining amount. *Dover Electronics and At/Comm.*

Reader Service No. 26

Create RS-485 networks with 496 nodes

A wiring concentrator for RS-422 and RS-485 networks lets users create RS-485 networks with up to 496 nodes and mix RS-422 and RS-485 systems on the same network. Model 290 uses an RS-232 master port and 16 slave ports that are independently programmable to either type of port. If each port is configured for RS-485 and considered a pseudo master port, users can expand the network to 496 nodes. Since separate driver/receiver circuits drive each of the 16 ports, a port failure is isolated from all other ports.

The 17W × 10D × 1.7H-in., aluminum-enclosed Model 290 can be changed from a standard desktop configuration to wall-mount or conventional 19-in. rack mounting. *Telebyte Technology; $725, delivery 2-4 weeks ARO.*

Reader Service No. 27

Apple supports SNMP

Apple Talk and TCP/IP network software now incorporate the Simple Network Management Protocol. System administrators can manage Macintosh personal computers on global networks using SNMP management consoles. Apple Talk Connection for Macintosh and TCP/IP Connection for Macintosh also provide a new System 7 service called the SNMP Manager that supports Watch Tower from Inter Con Systems Corp. and LAN Surveyor from Neon Software Inc. *Apple Computer; from $39 (single-user Apple Talk Connection), from $59 (single-user TCP/IP Connection).*

Reader Service No. 28

Modem communicates via memory-mapped scheme

The credit card-size, 2,400-baud Palm Modem card supports subnotebook and palmtop computers, communicating via a memory-mapped scheme, transmitting faxes, and running over 15 hours on two AA batteries. This PCMCIA Version 2.0 modem serves 8-bit processors such as the V20, Hopper Chip for the HP95LX, PC/Chip, V30, and the Zeo palmtop CPU. For the HP95LX, the modem contains a software interface compliant in format with Hewlett-Packard's system manager software. All of the software required to run the Palm Modem in the HP95LX is supplied on the card. *New Media Corporation; $259 (HP95LX version).*

Reader Service No. 29

Reader Interest Survey

Indicate your interest in this department by circling the appropriate number on the Reader Service Card.

Low 189 Medium 190 High 191

Product Summary

Joe Hootman
University of North Dakota

Manufacturer	Model	Comments	R.S.#
Boards			
Allen Systems	MP-11 SBC	Single-board computer designed for process control applications is based on the 8-bit 68HC11F1 microcontroller. The 4.5×5.5-in. MP-11 offers 16-MHz operation, power/ground planes for noise minimization, and a processor supervisory circuit. An expansion connector supports custom user circuitry or an optional A/D and D/A daughterboard. *$100 each (bare board/manual), $300 each (assembled/tested board); volume pricing available.*	80
Emulation Technology	HP-P5-PGA 14-UI preprocessor	Passive board including configuration software provides a timing analysis-only interface between Intel's Pentium microprocessor and most Hewlett-Packard logic analyzers. The preprocessor allows designers to make quick connections to a Pentium under test. The interface comes with built-in termination resistors. *$995 each; 10 days ARO.*	81
Gespac	GESMPU-46 SBC	Single-height Eurocard features a 20-MHz Cyrix 486 CPU chip, 486-code compatibility, two serial ports, and one bidirectional parallel printer port. Pairing with the GESVGA-1 enhanced VGA card produces AT compatibility in a form factor small enough for embedded industrial applications. *$1,795 each; available from stock.*	82
MNC International	MNC 1152 SBC	Single-board computer based on the 25-MHz Cyrix 486SLC processor promises a Landmark 2.0 CPU rating of 78 MHz. The passive backplane includes an SVGA CRT adapter, flat-panel (LCD and plasma) adapter, 1-Mbyte Flash memory, and clock/calendar. *$695 each, evaluation units; volume pricing available.*	83
Chips			
Cirrus Logic	CL-GD6440 LCD controller	Super VGA LCD device connects to a 32-bit local bus and a 32-bit video memory interface, offering desktop graphics capabilities in notebook computers. Two 256K×16 DRAMs provide 1 Mbyte of video memory, and integrated GUI assist functions support Microsoft Windows. The 208-pin QFP supports dual-scan color STN panels. *$40 each (5,000s).*	84
Mitsubishi Electronic Device Group	M38203M4/223M4/254M6 MCUs	Eight-bit microcontrollers with LCD controller and driver use 2.7V power. The ROM-based devices operate at up to 2 MHz with 2-µs minimum instruction executions and 8-mW typical power dissipation. *$4.85 to $6.75 each (10,000s).*	85

FOR DISPLAY ADVERTISING INFORMATION, CONTACT:

Southern California and Mountain States: Richard C. Faust, Douglas C. Faust, 24050 Madison Street, Suite 101, Torrance, California 90505; Tel: (310) 373-9604; Fax: (310) 373-8760.

Northern California and Pacific NW: William W. Hague, 9017 Peacock Hill Road, Gig Harbor, Washington 98332: Tel: (206) 858-7575; Fax (206) 858-7576.

East Coast: Gail A. Frank, Nancy Inserra, 82 Bethany Road, Suite 4, Hazlet, New Jersey 07730; Tel: (908) 264-1100; Fax: (908) 264-4340.

New England: Paul Gillespie, PO Box 6444, Holliston, Massachusetts 01746; Tel: (508) 429-8907; Fax (508) 429-8684.

Advertising Manager: Heidi Rex, 10662 Los Vaqueros Circle, Los Alamitos, California 90720-1264; Tel: (714) 821-8380; Fax: (714) 821-4010.

For production information, conference, and classified advertising, contact Marian Tibayan. *IEEE MICRO*, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, California 90720-1264; Tel: (714) 821-8380; Fax: (714) 821-4010; email: m.tibayan@compmail.com.

THE FOLLOWING INFORMATION IS AVAILABLE:

Contact the Publications Office;
to facilitate handling, please request by number.

- Membership application, student #203, others #202
- Periodicals subscription form for individuals #200
- Periodicals subscription form for organizations #199
- Publications catalog #201
- Compmail electronic mail brochure #194
- Technical committee list/application #197
- Chapters lists, start-up procedures #193
- Student scholarship information #192
- Volunteer leaders/staff directory #196
- IEEE senior member grade application #204
  (requires ten years practice and significant performance in five of those ten)

To check membership status or report a change of address, call the IEEE toll-free number, 1-800-678-4333. Direct all other Computer Society-related questions to the Publications Office.

PURPOSE

The IEEE Computer Society advances the theory and practice of computer science and engineering, promotes the exchange of technical information among 100,000 members worldwide, and provides a wide range of services to members and nonmembers.

MEMBERSHIP

Members receive the acclaimed monthly magazine *Computer*, discounts, and opportunities to serve (all activities are led by volunteer members). Membership is open to all IEEE members, affiliate society members, and others interested in the computer field.

IEEE COMPUTER SOCIETY®

A member society of the
Institute of Electrical and Electronics Engineers, Inc.

PUBLICATIONS AND ACTIVITIES

Computer. An authoritative, easy-to-read magazine containing tutorial and in-depth articles on topics across the computer field, plus news, conferences, calendar, interviews, and product reviews.

Periodicals. The society publishes eight magazines and seven research transactions. Refer to membership application or request information as noted at left.

Conference Proceedings, Tutorial Texts, Standards Documents. The Computer Society Press publishes more than 100 titles every year.

Standards Working Groups. More than 100 of these groups produce IEEE standards used throughout the industrial world.

Technical Committees. More than 30 TCs publish newsletters, provide interaction with peers in specialty areas, and directly influence standards, conferences, and education.

Conferences/Education. The society holds about 100 conferences each year and sponsors many educational activities, including computing science accreditation.

Chapters. Regular and student chapters worldwide provide the opportunity to interact with colleagues, hear technical experts, and serve the local professional community.

OMBUDSMAN

Members experiencing problems — magazine delivery, membership status, or unresolved complaints — may write to the ombudsman at the Publications Office.

EXECUTIVE COMMITTEE

President: James H. Aylor*
University of Virginia
Thornton Hall/Electrical Engineering
Charlottesville, VA 22903
Phone: (804) 924-6100
Fax: (804) 924-8818
E-mail: jha@virginia.edu

President-Elect: Laurel V. Kaleda*
Past President: Bruce D. Shriver*

VP, Technical Activities: Joseph Boykin (1st VP)*
VP, Conferences and Tutorials: Anneliese Von Mayrhauser (2nd VP)*
VP, Educational Activities: Gerald L. Engel†
VP, Membership Activities: Fiorenza C. Albert-Howard*
VP, Press Activities: Ronald G. Hoelzeman†
VP, Publications: Barry W. Johnson†
VP, Standards Activities: Gary S. Robinson†

Secretary: Mario R. Barbacci*
Treasurer: Michael Evangelist†
IEEE Division V Director: Bill D. Carroll†
IEEE Division VIII Director: V. Tom Rhyne†
Executive Director: T. Michael Elliott†

*voting member of the Board of Governors
†nonvoting member of the Board of Governors

BOARD OF GOVERNORS

Term Expiring 1993:
Fiorenza C. Albert-Howard, Jon T. Butler,
Michael C. Mulder, Yale N. Patt, Benjamin W. Wah,
Ronald Waxman, Akihiko Yamada

Term Expiring 1994:
Mario R. Barbacci, L. Felipe Cabrera, Wolfgang K. Giloi, Guylaine M. Pollock, John P. Riganati, Ronald D. Williams, Thomas W. Williams

Term Expiring 1995:
Fletcher J. Buckley, Doris L. Carver, Elliot J. Chikofsky,
Joanne E. DeGroat, Michael J. Flynn,
Mary Jane Irwin, Grace C.N. Wei

Next Board Meeting
February 26, 1993, 8:30 a.m.
Stanford Court Hotel, San Francisco, Calif.

SENIOR STAFF

Executive Director: T. Michael Elliott
Publisher: H. True Seaborn
Director, Conferences and Tutorials: Anne Marie Kelly
Director, Finance and Information Services: Deborah Rafal
Director, Board and Administrative Services: Violet S. Doan
Assistant to the Executive Director: Sandra K. Pfau

COMPUTER SOCIETY OFFICES

Headquarters Office
1730 Massachusetts Ave. NW
Washington, DC 20036-1903
Phone: (202) 371-0101
Fax: (202) 728-9614
E-mail: hq.ofc@compmail.com

Publications Office
10662 Los Vaqueros Cir.
PO Box 3014
Los Alamitos, CA 90720-1264
Membership and General Information:
(714) 8218380
membership@compmail.com
Publication Orders: (800) 272-6657
Fax: (714) 821-4010
E-mail: cs.books@compmail.com

European Office
13, Ave. de L'Aquilon
B-1200 Brussels, Belgium
Phone: 32 (2) 770-21-98
Fax: 32 (2) 770-85-05
E-mail: euro.ofc@compmail.com

Asia/Pacific Office
Ooshima Building
2-19-1 Minami-Aoyama, Minato-ku
Tokyo 107, Japan
Phone: 81 (3) 3408-3118
Fax: 81 (3) 3408-3553
E-mail: tokyo.ofc@compmail.com

IEEE OFFICERS

President: Martha Sloan
President-Elect: H. Troy Nagle
Past President: Merrill W. Buckley, Jr.
Secretary: Souguil J.M. Ann
Treasurer: Theodore W. Hissey, Jr
VP, Educational Activities: Edward A. Parrish
VP, Professional Activities: Charles K. Aledxander
VP, Publication Activities: Helen M. Wood
VP, Regional Activities: Luis T. Gandia
VP, Standards Activities: Wallace S. Read
VP, Technical Activities: Donald M. Bolle

IEEE MICRO

Editorial Calendar

FEBRUARY 1993

Automotive/traffic microelectronics

- Worldwide developments in microelectronics for traffic and driving assistance
- Improving traffic safety with electronics
- Latest developments from Japan, the European Prometheus, and US IVHS programs

Ad closing date: January 2

APRIL 1993

Advanced packaging and interconnection technology

- Critical packaging trends and issues
- Substrate and package technologies—for example, flexible, glass, or diamond substrates; few-chip or 3D packaging
- Attachment, bonding, and connection technologies, including fine-pitch surface mount, laser applications, known-good die, and interconnection trade-off analysis

Ad closing date: March 1

JUNE 1993

Hot Chips IV

- This extremely popular issue presents the latest developments in microprocessor and chip technology used to construct high-performance workstations and systems as presented at the annual IEEE Computer Society TCMM-sponsored symposium

Ad closing date: May 1

AUGUST 1993

Multitheme issue

- Multiprocessing
- Optical and biological computing
- Microcomputing to aid the handicapped
- Systems design
- DSP-based tools

Ad closing date: July 1

OCTOBER 1993

Far East issue

- Microcomputer and RAM technology
- High-performance parallel computer research
- Current TRON Project offerings, Japan's standard computer system
- Singapore technology update

Ad closing date: September 1

DECEMBER 1993

Special standards issue

- Buses and interconnections
- Processor architectures
- Current activities: IEEE and other standards bodies
- How standards can help designers

Ad closing date: November 1

IEEE Micro helps designers and users of microprocessor and microcomputer systems explore the latest technologies to achieve business and research objectives. Feature articles in *IEEE Micro* reflect original works relating to the design, performance, or application of microprocessors and microcomputers.
All manuscripts are subject to a peer-review process consistent with professional-level technical publications. *IEEE Micro* is a bimonthly publication of the IEEE Computer Society.

Advertising information: Contact Marian Tibayan, Advertising Department, IEEE Computer Society, PO Box 3014, Los Alamitos, CA 90720-1264; (714) 821-8380; fax (714) 821-4010.

Articles may change. Please contact editor to confirm.